A COMPREHENSIVE EMPIRICAL INVESTIGATION ON ANOVA EFFECT SIZES

A Dissertation

by

YUANYUAN ZHOU

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Bruce Thompson
Committee Members,	Victor Willson
	Mary Margaret Capraro
	Oi-man Kwok
	Susan Skidmore
Head of Department,	Victor Willson

December 2015

Major Subject: Educational Psychology

Copyright 2015 Yuanyuan Zhou

ABSTRACT

The present journal article formatted dissertation was a comprehensive investigation of ANOVA effect sizes. In the first study, the author examined the extent to which ANOVA practices have changed in comparison with a methodology review conducted 15 years ago, which include the examination of validity assumptions, sample sizes, and effect size indices. The author reviewed all articles published in 2012 in three APA journals (*Journal of Applied Psychology (JAP), Journal of Counseling Psychology (JCP),* and *Journal of Personality and Social Psychology (JPSP).* Results indicated that the use of ANOVA is proportionally less than previously indicated, but still very popular in practice. Researchers still rarely verify whether the validity assumptions are satisfied, but reporting effect size statistics is on the increase.

In the second study, the author examined the accuracy and robustness of estimates of practical significance (i.e., $\hat{\eta}^2$, partial $\hat{\eta}^2$, $\hat{\varepsilon}^2$, partial $\hat{\varepsilon}^2$, $\hat{\omega}^2$, and partial $\hat{\omega}^2$) in a 2 × 3 two-way fixed-effects ANOVA. The study extended the exploration of these effect sizes in the presence of assumption violations and is generalized to the more common case of multi-factor ANOVAs. The results revealed that: the classical forms were more stable; $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ were not always better estimates than $\hat{\eta}^2$; sample sizes, group size ratio, heterogeneity of variance, population effect sizes, pairings, and degrees of freedom all affected the effect sizes estimate.

In the third study, the author examined the accuracy and robustness of estimates of Intraclass Correlation Coefficient (ICC) in a 2 × 3 two-way mixed ANOVA. Results

indicated that the accuracy and robustness of estimation were mainly affected by two components: sampling error due to random-effects and sampling error due to random sampling of a sample. ICC estimates are robust across different studies as long as the number of levels for the random effect is the same. Researchers should be cautious to utilize the ICCs' estimates when the design differs from the design investigated here.

DEDICATION

In memory of Lingying Zhou, the greatest mom. You always supported me through everything and believed that I can be the one I want to be. You left fingerprints of grace on my life. You will not be forgotten.

ACKNOWLEDGEMENTS

I want to acknowledge many people that helped me to finish my dissertation. The one that I would like to express my deepest gratitude is my committee chair, Dr. Bruce Thompson. He gave me very professional guidance, freedom, and patience. He influenced me not only because he is erudite, but also because he is rigorous in research and life. He is the best professor I have ever met that cares about students and devotes himself to helping students succeed.

I would also like to thank my other committee members. Dr. Susan Skidmore is both my committee member and a close friend. I received detailed guidance from her through every step. She encouraged me and consoled me during my difficulties. Dr. Oi-man Kwok, Dr. Victor Willson, and Dr. Mary Margaret Capraro have also given their support and guidance throughout my doctoral studies.

I also want to extend my gratitude to my friends and other department faculty and staff. They filled in my memories at Texas A&M University; especially thanks my friends Kevin and Jenifer. They helped me taking care of my two little ones and saved more time for me to do my research.

Finally, thanks to my husband, my parents, and my mother-in-law. Without their support, I would never have been able to finish my dissertation.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

INTRODUCTION

Analysis of variance (ANOVA) *F* tests have been identified in past methodology reviews as the most popular data analytical technique in educational research (Keselman et al., 1998; Skidmore & Thompson, 2010). Based on whether or not the levels of ways are treated as a sample from a larger population, ANOVA *F* tests can be categorized into three classes: fixed-effects ANOVA (i.e., the levels of ways are all enumerated and do not generalize to a larger population), random-effects ANOVA (i.e., the levels of ways are treated as a sample from a larger population and thus are generalizable to a larger population), and mixed-effects ANOVA (i.e., the ANOVA contains both fixed effects and random effects). ANOVA *F* tests are often used to compare group means for statistical significance, especially when three or more means are compared (the *t*-test is the alternative choice for comparing two group means). ANOVAs are also applied in experimental field trials to test whether a difference exists between different conditions. Among the three types of ANOVA models, the fixed-effects model is most frequently used in experimental designs, in which measured treatments are the only research interests. But in some cases, the measured variables (such as time or personal traits) need to be generalizable to a broader situation; in such cases, the mixed-effects or random-effects models may be more appropriate (Hedeker & Gibbons, 2006). In addition, the random-effects ANOVA is the theoretical foundation for Generalizability theory and is widely applied in psychometric variance partitioning (Kane, 2002; Shavelson & Webb, 1991). For the same reason, the random-effects model is also called a variance components model. Finally, the random-effects model is also a special case of the

hierarchical linear model in which the dataset demonstrates hierarchical structure, and the difference in the lower level is associated with the hierarchy (i.e., random effects) (Hox, 2002).

However, among those researchers who applied ANOVA *F* tests in their research, only "few researchers read ANOVA theory before starting to analyse" (Cardinal & Aitken, 2013, p. 1). It may be a stringent requirement to demand that all quantitative researchers understand the math and statistics undergirding the ANOVA *F* tests, but they should at least know that ANOVA is not an omnipotent data-analytical technique that can be applied to any circumstance. They should know that ANOVA is valid only when the three main assumptions (i.e., independence of observations, homogeneity of variance, and normally distributed residuals) are reasonably satisfied. However, a methodology review conducted 70 years after the ANOVA was originally conceptualized by Ronald Aylmer Fisher in the 1920s (Fisher, 1925) revealed that researchers "rarely verify that validity assumptions are satisfied and … typically use analyses that are nonrobust to assumption violations" (Keselman et al., 1998, p. 350). In fact, the criticism of overuse, misuse, and misinterpretation of ANOVA in behavioral science has never ceased. Neglecting to verify the validity assumptions is only a portion of the misuse, beyond that, there are at least three other malpractices of ANOVA usage: (1) using a fixed-effects model when random-effects or mixed-effects models were more appropriate (Bennington & Thayne, 1994); (2) artificially grouping continuous variables into categories so as to fit the ANOVA tests that require the predictors be in the nominal scale and thus cause the loss of useful information (Hester, 2000); and (3)

"*unconsciously* and *erroneously*" making causal inferences simply because an ANOVA test was conducted (Thompson, 2006, p. 386). A major reason that causes the malpractice of ANOVA is the lack of understanding of ANOVA concepts and limitations (e.g., theoretical assumption for fix-effects and random-effects models, clarification of the inference of research results, and the difference between experimental design and ANOVA test).

Though ANOVA has been identified as the most popular data-analytic technique, historical methodology reviews have also discovered a decreasing trend of ANOVA use in the social sciences (Bangert & Baumberger, 2005; Baumberger & Bangert, 1996; Edgington, 1964; Edgington, 1974; Goodwin & Goodwin, 1985; Kieffer, Reese, & Thompson, 2001; Willson, 1980). Skidmore and Thompson (2010) reviewed ten statistical techniques reviews of some major educational and psychological journals ranging from 1948 to 2001, and revealed that in educational research, the use of ANOVA had a "marked decrease from the 1970s to the 1990s" (p. 781); and in psychological research, the "ANOVA techniques have a definite curvilinear relationship [with time] and appear to have been steadily decreasing in usage beginning in the 1990s" (p. 785). Two reasons account for the steady decrease of ANOVA usage. One was that more and more researchers gradually became conscious of the misuse of ANOVA as their understanding of ANOVA concepts increased. Thus, they were more likely to avoid overuse and misuse of ANOVA. Similarly, newer statistical methods have been developed and were applied on datasets that would have been analyzed with ANOVA if the research had been conducted decades ago. For example, with the development of

multilevel modeling, the continuous variable no longer needs to be categorized into nominal scale thus protecting the unnecessary loss of any useful information (Paterson & Goldstein, 1991). In addition, structural equation modeling (SEM) is increasingly applied in the behavioral sciences and has the potential to inform causal inferences (Thompson, Diamond, McWilliam, Snyder, & Snyder, 2005). However, with the development of new statistical techniques, a new erroneous trend in social science is to incorporate as many variables as possible, as if the more variables are measured, the higher the quality of the research. As Cohen commented (1990), "I have encountered too many studies with prodigious numbers of dependent variables, or with what seemed to me far too many independent variables, or (heaven help us) both" (p. 1304). Another erroneous trend is the overuse or misuse of new developed statistical techniques when simple traditional statistical methods were sufficient for the research purpose. It is one thing to employ a new technique when the technique is more appropriate than a more traditional approach. However, it is something quite different to use a more complicated technique when the traditional approach would suffice. If quantitative researchers in social sciences already have difficulty in mastering basic statistical methods, what guarantee is there that a more complicated statistical technique could be applied correctly? Therefore, Cohen (1990) summarized the principles he learned as a methodologist as "less is more" and "simple is better" (p. 1304).

The Null Hypothesis Significance Testing and Effect Sizes

Historically, null hypothesis significance testing (NHST) has dominated quantitative research. However, in the meantime, NHST has been the object of

controversy among social scientists who favor and those who oppose. The misuse and misinterpretation of NHST (e.g., $p < .05$ is equivalent to importance, p measures results replicability, and NHST as a vehicle to avoid judgment) has raised major methodological concerns in research (Daniel, 1998; Thompson, 1999). Criticisms on NHST rose rapidly in recent decades from various areas (e.g., education (Thompson, 1996), psychology (Cohen, 1994; Hagen, 1997; Schmidt, 1996), political science (Gill, 1999), and biology (Anderson, Burnham, & Thompson, 2000)). NHST brings the tautology to researchers (i.e., researchers have collected a certain amount of samples but need to conduct a statistical test to evaluate whether the sample size is sufficient) and is harmful for the accumulation of knowledge; NHST invokes a nonsensical comparison because a sufficiently large sample always leading to a statistically significant result; and NHST also brings an inescapable dilemma for researchers because they want smaller samples so as to fail to reject the preliminary methodological assumption hypotheses, but a larger sample to reject the substantive research hypotheses (Thompson, 1993). Distinguished scholars undertook debates on whether or not statistical significance tests should be banned. In 1999, the APA Task Force issued its recommendations: APA did not recommend banning the use of statistical significance testing, but, they did *strongly recommend* the use of effect sizes and confidence intervals (American Psychological Association, 2001). And in the most recent APA (2010) manual, APA further emphasized that the NHST "is but a starting point", *effect sizes* and other additional elements "are needed to convey the most complete meaning of results" (p. 33).

The emphasis on effect sizes has been rapidly rising in the past decades (American Psychological Association, 2001, 2010). NHST uses the *p*-value to measure whether or not the null hypothesis is true. However, NHST will not bring researchers one step closer to the true magnitude of difference, no matter how many NHSTs were conducted. NHST does not contribute much to the accumulation of knowledge and has nothing to do with thinking meta-analytically about research. An effect size, however, "measures the degree to which such null hypothesis is wrong" (Grissom & Kim, 2012, p. 5), which quantifies the deviation of the sample statistic to the null hypothesis. Effect sizes are essential to the accumulation of knowledge and also enable thinking meta-analytically about of research. As Thompson (1999) claimed, "Most single studies are important primarily only as building blocks within a cumulative body of evidence" (p. 170). Effect size estimates the true difference makes this possible to think retrospectively on research. Today, "at least 24 journals in various fields *require* [emphasis added] that authors of research reports provide estimates of effect size" (Grissom & Kim, 2012, p. xiii), and the APA manual (2001) emphasized that "failure to report effect sizes" is one "kinds of *defects* [emphasis added] in the design and reporting of research" (p. 5).

Effect Sizes for ANOVA

With the increase emphasis on reporting effect size, the new problem has emerged that "authors do not recognize effect sizes produced in their own analysis" (Skidmore, 2009, p. 4) or do not know which forms of effect sizes tend to provide the least biased and most robust estimates. And "applied researchers have continued to rely almost exclusively on [certain]… indicators of effect when reporting their findings"

(Olejnik & Algina, 2000, p. 241), which may or may not be the best indicator to measure the effect.

There are many forms of indicators to measure ANOVA effect sizes. Based on a number of dimensions, effect sizes can be categorized as score-world effect sizes (also called as standardized effect sizes), which are in the unsquared metric; and area-world effect sizes (also called variance-accounted-for effect sizes), which are in the squared metric (Fidler & Thompson, 2001; Thompson, 2006). For the one-way fixed-effects ANOVA, effect sizes can be either in a score-world scale (e.g., Cohen's *d*) or area-world scale (e.g., η^2, ε^2, and ω^2). But in more general cases (i.e., multi-way ANOVA), the area-world effect sizes that measure "the proportion of variance of the scores on the dependent variable that is related to variation of the independent variable" (Grissom & Kim, 2012, p. 207) are more appropriate, because in multi-way ANOVA, the standard deviation used to standardize the mean difference in one way is affected by the other ways and interaction. Based on whether or not the total variances (i.e., the denominator) include a portion of factor variance that is not from the target factor, the effect sizes can be categorized into classical forms that include all variances (e.g., $\eta^2 = \frac{SS_{effect}}{SS_{total}}$), and alternative forms that exclude other non-target factor variances (e.g., $\eta_p^2 = \frac{df_A F_A}{df_A F_A + df_E}$) (Cohen, 1973, p. 107).

Also, based on whether or not the ANOVA has random-effects, the estimate of practical significance can be categorized as a fixed-effects model effect size or a

random-effects model (including the mixed-effects model) effect size. Most of the time effect sizes appearing in journals or textbooks measure the fixed-effects ANOVA effect sizes. Grissom and Kim (2012) in their popular effect sizes book clarified that all discussions that discuss effect size for ANOVA assume the fixed-effects model. And Hedges (1983) commented, "Statistical theory proposed previously has stressed the estimation of fixed but unknown population effect sizes" (p. 388). Indeed, discussions of random-effects ANOVA effect sizes are not as frequent as fixed-effects ANOVA effect sizes.

Intraclass correlation coefficients (ICC) are among the most popular random-effects ANOVA effect sizes. ICCs "are commonly used in behavioral measurement, psychometrics, and behavioral genetics" (McGraw, & Wong, 1996, p. 30). And the most recognized form of ICC is as the reliability coefficient. There are various ICC reliability coefficients. Based on the categorization proposed by Shrout and Fleiss (1979), there are six types of ICCs. To determine which type of ICC is most appropriate three questions may be asked: (1) is it a one- or two-way analysis of variance? (2) Can one effect be ignored in the reliability index? And (3) what is the unit of reliability? Mcgraw and Wong (1996) added two more questions for choosing the correct ICC: (4) is it the reliability of absolute agreement or the reliability of consistency, and (5) does it include interaction? Based on Mcgraw and Wong's definition, the types of ICC reached 10.

Organization of Document

This dissertation includes three completed studies, all drafted as the manuscripts planned for publication in peer-reviewed journals. The first study presents a systematic

review of current ANOVA practices in three APA journals. The review addresses the following three questions: (1) is ANOVA still a frequently used data-analytic technique? (2) How are ANOVAs applied in educational and psychological research? And (3) are there improvements on assumptions verifications and reporting of effect sizes, compared with ANOVA practices 17 years ago.

The second study presents a simulation study on six frequently used effect sizes (i.e., η^2, ε^2, ω^2, η_p^2, ε_p^2, and ω_p^2) based on a 2 × 3 fixed-effects ANOVA. This simulation considers 4 × 3 × 2 × 3 = 72 conditions with 72 × 10,000 = 720,000 total replications. The 72 conditions were: four Cohen's *f*s (0, 0.1, 0.25, and 0.4), three variance ratios (1:1 for the A-way and 1:1:1 for the B-way, 1:1.5 for the A-way and 1:1:1.5 for the B-way, and 1:2 for the A-way and 1:1:2 for the B-way), three types of average cell sizes (6 and 36), three types of pairings (balanced, positive pairing, negative pairing). The second study addressed the following four questions: (1) which one is better: classical forms or partial alternative forms? (2) Which one is better: $\hat{\eta}^2$, $\hat{\varepsilon}^2$, and partial $\hat{\omega}^2$? (3) What conditions most affected the effect size estimates? And (4) how does number of levels affect the effect size estimates?

The third study presents a simulation study on estimates of ICCs based on a 2 × 3 mixed-effect ANOVA model. ICC as a type of effect sizes, the estimation of its quantity should also be affected under different conditions. This study explores how heterogeneity, unequal group size, and the combination effect of heterogeneity and

unequal group size, size of the fixed-effect, total sample sizes, and whether or not there is an interaction affect the accuracy and robustness of ICC estimation.

As a whole, the dissertation provides a comprehensive investigation on ANOVA (both fixed-effects and random-effects models) practices and the accuracy and robustness of estimation of ANOVA effect sizes under various conditions.

RESEARCHER PRACTICES AND THE TENABILITY OF STUDY RESULTS: A REVIEW OF ANOVA PRACTICES IN THREE APA JOURNALS

The prevalence of ANOVA in the educational and psychological literature has been well-documented (Kieffer, Reese, & Thompson, 2001; Skidmore & Thompson, 2010). ANOVA is useful in substantive studies to analyze mean differences across *k* groups. ANOVA is also widely used in psychometric variance partitioning (Kane, 2002; Shavelson & Webb, 1991). Like all parametric analyzes, the conclusions drawn from ANOVA results are dependent upon the extent to which statistical assumptions are met. Numerous works have reported the lack of robustness of the *F* ratio (and resulting $p_{\text{calculated}}$) in the presence of an unbalanced design and heterogeneous variance (Glass, Peckham, & Sanders, 1972; Harwell, Rubinstein, Hayes, & Olds, 1992; Lix, Keselman, & Keselman, 1996). Monte Carlo evidence has demonstrated the negative impact heterogeneity of variance can also have on estimates of practical significance in one-way ANOVA designs (Keselman, 1975; Skidmore & Thompson, 2013). Of course, the assumption of homogeneity of variance is never fully met in applied research. A practical question for researchers is "whether the plausible violations of the assumptions have serious consequences on the validity of probability statements based on the standard assumptions" (Glass et al., 1972, p. 237).

Current reform efforts are less focused on the *p*-value and more focused on estimates of practical significance (i.e., effect sizes) and the corresponding confidence intervals (Kelley & Preacher, 2012; Thompson, 2002). The American Psychological Association (APA) Publication Manual (2010) noted that

> Historically, researchers in psychology have relied heavily on null hypothesis statistical significance testing (NHST) as a starting point for many (but not all) of its analytical approaches. APA stresses that NHST is but a starting point and that additional reporting elements such as effect sizes, confidence intervals and extensive description are needed to convey the most complete meaning of results. (p. 33)

Methodological research reviews are frequently used to identify the trends in quantitative research practice. In brief, what are the typical data-analytic techniques used in applied journals within a certain period of time? Such reviews of practice are important in that "journals both create and mirror their fields" (Silverman, 1987, p. 40). However, methodological research reviews consistently find a substantial gap between recommended inferential methods and the methods actually adopted by applied researchers. For instance, Keselman et al. (1998) presented a review of researchers' ANOVA practices in prominent journals from 1994 and 1995 including validity assumptions, sample sizes, and effect size indices. The review results indicated that researchers (1) rarely verified that the ANOVA distributional assumptions were satisfied, (2) typically used regular ANOVA tests that were not robust to assumption violations, (3) rarely reported effect size statistics, and (4) rarely performed power analyses to determine the sample size requirements. Understanding researcher practices provides an opportunity to make recommendations about best practices, offer guidance on graduate training, and provide a basis for what statistical knowledge is needed to read, engage in, and contribute to a field.

Purpose

Although methodological research reviews of the data analytic techniques have a long history in educational and psychological fields (Edgington, 1964, 1974; Goodwin & Goodwin, 1985; Kieffer et al., 2001; Willson, 1980), new methodological reviews are still needed to capture emerging trends and the movement of research field. Today, 17 years after the Keselman et al. (1998) review I investigate the extent to which ANOVA practices have changed in light of current reform efforts. I reviewed ANOVA practices in the *Journal of Applied Psychology (JAP), Journal of Counseling Psychology (JCP), Journal of Personality and Social Psychology (JPSP)*. My primary research question was, how prevalent is the use of ANOVA in comparison to historical practices (i.e., in the period Keselman et al. [1998] review was conducted) given that the use of new techniques are on the increase every year? My second research question was to what extent have ANOVA practices changed after recommendations were clearly offered by Keselman et al. (1998)? The third research question was what changes are still needed based on current ANOVA practices?

Method

Criteria for Study Inclusion

Three *APA* journals (i.e., *JAP*, *JCP*, and *JPSP*) were chosen as the target journals. All articles published in 2012 were collected. A total of 87 entries were located for *JAP*, 61 for *JCP*, and 147 for *JPSP*. The following keywords were used for a

preliminary electronic review for evidence of a means test and the identification of inappropriate articles:

> ANOVA, OVA, Factorial, Non-Factorial, *F* test (*F*-test), Omnibus test, One way (One-Way), Two Way (Two-Way), Multi-Way, Brown-Forsythe, Welch, Mann-Whitney *U*, Kruskal-Wallis, Friedman, *t* test (*t*-test), ANCOVA, Means test, James, Post hoc.

Articles that did not contain any of the keywords were excluded from further review. This procedure resulted in a total of 224 articles (*JAP*: 68, *JCP*: 39, and *JPSP*: 117). The manual review further excluded 6 articles: 5 of which were qualitative research articles and 1 article was retracted due to academic fraud. The remaining 218 articles were subjected to a manual coding process. A coding scheme was developed and transferred to an online questionnaire. Six variables were coded for the 218 articles that underwent a manual coding process: (1) types of means tests; (2) types of *F*-tests; (3) report of textbooks or article references to justify the use of ANOVA; (4) report of statistical packages; (5) use of the terms of "small", "medium", or "large" to qualify effect sizes; and (6) number of other analysis of variance means tests excluding fix-effects ANOVA (i.e., MANOVA, MANCOVA, mixed-effects ANOVA, and repeated measures, etc.). Each element of the coding scheme is discussed in greater detail in the results section. Eighty-two (82) articles that reported the use of between-subject fixed-effects ANOVA *F* tests, which are the most popular data-analytic technique among all analysis of variance *F* tests, underwent the full coding process.

Table 1

Coding Scheme for Online Questionnaire

Coding questions	Options
General Review	
Types of mean test	*t* test, *F* test, Brown and Forsythe, Welch, Mann-Whiney *U*, Kruskal-Wallis, Friedman, Planned Contrasts, others
types of *F* test	Between-subject ANOVA(independent *t*-test), ANCOVA, MANOVA, Mixed ANOVA(repeated measures, paired *t*-test), others
Textbook or article references to justify the use of ANOVA	Not given, others
Report of statistical packages	none was given, SPSS, STATA, SAS, R, Other
Use the terms of "small", "medium", or "large" to quantify effect sizes	None; Cohen (1988) Statistical power analysis for the behavioral sciences (test); Cohen (1990) Things learned so far Amer. Psych.; Cohen (1992) A power primer Psych. Bulletin; Cohen (1992) Statistical power analysis current directions Psych. Science; Cohen (1994) Earth is round Amer. Psych.; Others.
Number of uncoded *F* tests research techniques	MANOVA, Mixed ANOVA, Repeated measures, paired *t*-test, Random-effect ANOVA
Full Review	
Number of between-subject fixed-effects ANOVAs	1, 2, 3, 4, 5, more than 5
Report of statistical violation assumptions	None, independence of observations, variance homogeneity, distribution (normality)
Report of methods to deal with assumption violations	Nothing because assumptions were not mentioned; Nothing because assumptions were not violated; Transformation; Use of nonparametric analyzes; Winsorizing & Trimming; Converting continuous variables to categorical variables.
Report of method to test for violation assumptions	None, Levene's, Shapiro-Wilks, Bartlett's test, Test was run, but no name was given
Indicate the number of ways, levels, and design type in ANOVA.	
Report of post hoc tests	None, LSD, Bonferroni, Sidak, Scheffe, Tukey, Duncan, Hochberg, Gabriel, Walter-Duncan, Dunnet, Others
Ratio of the largest to smallest standard deviation	
Ratio of the largest to smallest group size	
Sample sizes	
The way p values were reported	$p < .05$, $p < .01$, $p < .001$, $p > .###$, $p = .###$, ns, others
Reported effect sizes	None, η^2, partial η^2, ξ^2, ω^2, *d*, *f*, *r*, other

In the full coding process, another 13 characteristics were coded: (1) number of between-subject ANOVA *F*-tests; (2) report of statistical violation assumptions; (3) report of method to test for violation assumptions; (4) report of methods to deal with assumption violations; (5) number of ways; (6) number of levels for each way; (7) design type (factorial or non-factorial); (8) report of post hoc tests; (9) ratio of the largest to smallest standard deviation; (10) ratio of the largest to smallest group size; (11) sample sizes; (12) the way *p* values were reported; and (13) reported effect sizes. See Table 1 for details.

The first step of the coding process was to exclude those articles that did not include any keywords. However, an article that contained a keyword during the electrical review process did not necessarily indeed contain *F* tests. For example, an article that contained the keywords "Friedman" did not necessarily have the Friedman test. "Friedman" might either be the author's name for this article or the author for referenced articles. Therefore, the number of articles containing any *F* tests is indeed smaller than 224. In addition, MANOVAs, between-subject random-effects ANOVAs and repeated measures (including mixed-effect ANOVA and paired *t*-test) were excluded from the full coding process. Finally, because some articles contained multiple studies per article and to maintain consistency in the coding process the decision was made to code the first five ANOVAs within the first study when multiple studies were reported per article. Therefore, if the study or the first study contained more than five ANOVAs, only the first five ANOVAs were coded. This decision underestimated the proportion of ANOVA *F* tests used in the three journals because I did not code ANOVA

F tests reported other than the first study and I did not code the remaining ANOVA *F* tests when the first five ANOVA *F* tests were coded. Consequently, the resulting total number of articles that contained at least one between-subject fixed-effects ANOVA and thus were subjected to the full coding process was 82. Figure 1 presents the article inclusion criteria and the article review process. One thing to note is that all 218 articles that had means tests underwent manual reviews.

Interrater Reliability

To guarantee consistency in the review process, the coding was completed twice. During the first round of coding, I met with my committee member, Dr. Susan Skidmore, every other week to discuss issues that I encountered during the process of coding to make sure that the coding was consistent. Susan also randomly coded 10% of the articles to establish interrater reliability. Finally, I reviewed all of the articles a second time to reduce any possible errors that may have occurred during the first coding process. I compared the first and second coding results, when agreement in both coding process was not reached, I consulted Dr. Susan Skidmore for advice until a consensus was reached.

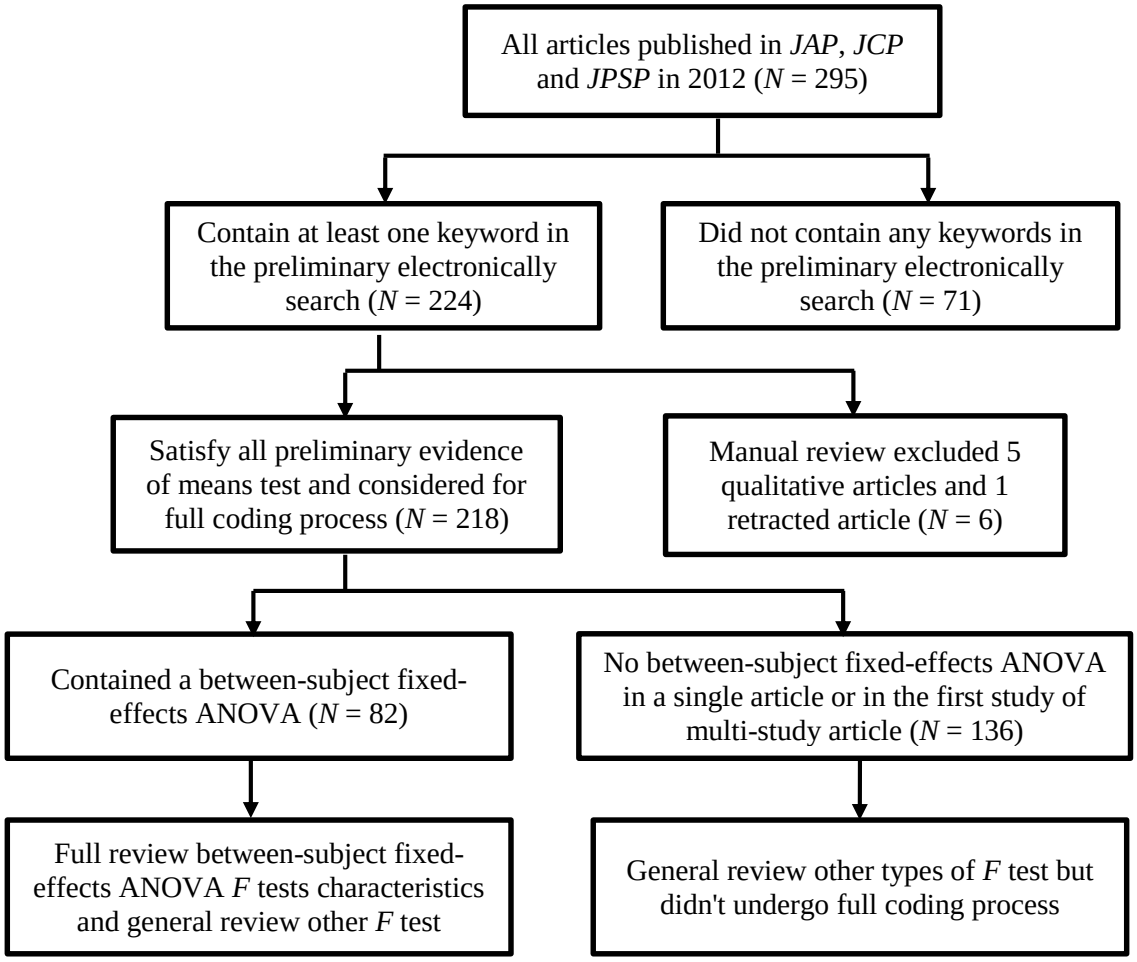

Figure 1. Article inclusion criteria decision sequence.

Results

General Review

Types of ANOVA *F* tests. As shown in the Table 2, among the 218 articles that underwent manual review, 53% (n = 116) articles reported analysis of variance F/t tests, of which 82 articles contained between-subject fixed-effects ANOVA F tests, 9 articles contained between-subject random-effects ANOVA F tests, 40 articles contained within-subject ANOVA F tests (including mixed-effects ANOVAs and repeated measures), and 15 articles contained MANOVA/MANCOVA F tests. Because many articles reported more than one type of analysis of variance F test (for instance, an article might have contained both the between-subject fixed-effects ANOVA F tests and the between-subject random-effects ANOVA F tests), the sum of all types of F test research techniques is greater than the total number of articles that reported analysis of variance tests.

Between-subject fixed-effects ANOVA F tests were the most popular F tests among all analysis of variance techniques. The between-subject random-effects ANOVA F tests were also fairly common, as 22. 48% of the total reviewed articles (4.13% for between-subject random effects ANOVA, and 18.35% mixed-effects/repeated measures ANOVAs) contained analysis of variance F tests that treated at least one way as random. Researchers showed different preferences for different analysis of variance techniques in the three journals: in *JAP*, to estimate the homogeneity within groups, one-way random ANOVA F tests were used to calculate the intra-class correlation coefficient; while in *JPSP*, because personal traits are often the major research focus, mixed ANOVA designs

or repeated measures were frequently used as the analytical techniques. On the other hand, MANOVA/ MANCOVA were equally uncommon across the three journals; neither was often used.

Table 2

Journal Source and Frequency of OVA Reported

Journal / Statistic	Between-subject fixed-effects ANOVA / independent *t*-test	Between-subject random-effects ANOVA	Mixed-effects ANOVA / repeated measures / paired *t*-test	MANOVA / MANCOVA
JAP	20	9	6	6
JCP	16	0	3	5
JPSP	46	0	31	4
Total	82	9	40	15
Percentage	37.61%	4.13%	18.35%	6.88%

Note. Percentage reflects the $N = 218$ preliminarily reviewed quantitative articles. The sum of the percentage not equal to 100% because: (a) not all of the 218 articles actually contain analysis of variance *F*/*t* test, and (b) some articles reported more than one type of *F*/*t* test.

Statistical citations. Among the 218 articles that underwent manual review, only six articles cited references to justify the use of statistical techniques in the method section. Provided in Table 3 are the six articles and the reason given for citing a reference: three articles (i.e., *JAP*4-11, *JCP*3-15, and *JPSP*7-11) used these references to justify the use of alternative methods when ANOVA assumptions were violated (e.g., heterogeneity of variance and non-normality) and uneven sample sizes existed. I applaud the three articles' authors because during my full review, I found very few authors noted the possible violation of ANOVA assumptions. Indeed, most authors did not even cite

references to justify the use of alternative methods. For the other three articles: one (i.e., *JPSP*3-5) cited a reference as the criterion to select a candidate covariate; one article (i.e., *JAP*5-9) cited a reference to determine the validity and reliability of the unit goal orientation variables, and one (i.e., *JAP6*-12) used the reference to determine if a hierarchical data set could be aggregated into a team level.

Table 3

Articles Use of References to Justify the Chosen Analytical Techniques

Code	Reference	Reason
*JAP*4-11	Wilcox, 2005[a]	Alternative method due to concerns with covariance heterogeneity, uneven sample sizes and non-normality
*JAP*5-9	Bliese, 2000[b]	Determine the validity and reliability of the unit goal orientation variables
*JAP*6-12	Bliese, 2000[b]	Determine if hierarchical data set can be aggregated into team level
*JCP*3-15	J. Cohen, Cohen, West, & Aiken, 2003[c]	Method to deal with data that depart from normality
*JPSP*3-5	Darlington, 1996[d], Walton, & Cohen, 2007[e]	Criterion to select candidate covariate
*JPSP*7-11	Erceg- Hurn & Mirosevich, 2008[f]	Alternative method due to concerns with non-normality

Note. *JAP*4-11 means *Journal of Applied Psychology*, issue 4, the 11th article, etc.

a. Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego, CA: Academic Press.
b. Bliese, P. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research and methods in organizations* (pp. 512–556). San Francisco, CA: Jossey-Bass.
c. Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral science* (3rd ed.). Mahwah, NJ: Erlbaum.
d. Darlington, R. (1996). *How many covariates to use in randomized experiments?* Retrieved from http://www.psych.cornell.edu/darlington/covarnum.htm
e. Walton, G. M., & Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. Journal of Personality and Social Psychology. *Journal of Personality and Social Psychology*, 92, 82–96. doi:10.1037/0022-3514.92.1.82
f. Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, 63, 591– 601. doi:10.1037/0003-066X.63.7.591

Terms to quantify effect sizes. Even though Cohen cautioned against the thoughtless use of "small", "medium", and "large" to quantify effect sizes when there was existing literature that could more precisely describe what a "small", "medium", or "large" effect was in a particular discipline (Cohen, 1988, p. 532), the problematic terms still occasionally appear in published articles. Among the 218 initially reviewed articles, 23 articles used these terms, and 10 articles (see Table 4 for details) even provided references (e.g., Cohen, 1988, 1992) to support the use of these terms even though Cohen himself recommended that his benchmarks not be used.

As shown in Table 4, only one article used the term appropriately to set up the simulation conditions. In the other nine articles, the terms of "small", "medium", and "large" were erroneously treated as standards to quantify the measured effects, as if "small" equated to a non-essential effect, "medium" equated to a moderate effect, and "large" equated to an important effect. But the appropriate approach to interpreting effect sizes is not simply to ascribe a predetermined qualifier. A small value of effect, like $d = 0.001$ may be very important in certain fields, but may be negligible in other research areas (Thompson, 2006).

Software packages. Reporting the statistical software used for data analysis was not common in the three journals examined in the present study. Among the 218 reviewed articles, 80% did not mention anything about the package used for data analysis. Among those articles that reported the name of the software, SPSS was noted most frequently (22 articles), followed by SAS (12 articles), HLM (5 articles), MPLUS (4 articles), LISREL (3 articles), R (1 article), and STATA (1 article).

Table 4

Articles that Used the Terms "Small", "Median", and "Large" to Describe Effect Sizes

Code	Reference	Reason
*JAP*2-2, *JCP*1-11, *JCP*2-12, *JPSP*8-7	Cohen, 1992[b]	To justify that the effect size obtained from study is a small effect size
*JCP*1-12, *JCP*2-6, *JCP*2-8	Cohen, 1992[b]	To justify that the effect size obtained from study is a median effect size
*JPSP*8-1	Cohen, 1988[a]	To justify that the effect size obtained from study is a large effect size
*JCP*3-13	Sink & Stroh, 2006[c]	The rule of thumbs to interpret effect size
*JAP*5-3	Cohen, 1988[a]	Recommended condition for simulation design

Note. *JAP*2-2 means Journal of Applied Psychology, issue 2, the second article, etc.

a. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
b. Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159. doi:10.1037/0033-2909.112.1.155
c. Sink, C., & Stroh, H. (2006). Practical significance: The use of effect sizes in school counseling research. *Professional School Counseling*, 9, 401–411.

Full Review

The proportion of different means tests. Among the 82 articles that contained between-subject fixed-effects ANOVAs a total of 261 means tests were documented, of which, 138 (52.9%) were traditional ANOVA *F* tests, 108 (41.4%) were independent *t* tests, 7 (2.7%) were ANCOVA *F* tests, 2 were Welch alternative ANOVA *F* tests, 5 were planned contrast tests, and 1 was a nonparametric bootstrapping analysis test (see Table 5 for details). Traditional ANOVA *F* tests and *t* tests were 94% of all documented means tests. This large percentage forces me to question if all 94% of the educational

and psychological researches met the ANOVA prerequisites: independence, homogeneity of variance, and normality, because ANOVA is a valid statistical method only when all three assumptions are reasonably satisfied.

Table 5

The Proportion of Various Means Tests in JAP, JCP and JPSP

Journal	*t*-test	ANOVA	ANCOVA	Welch test	Planned Contrasts	Other
JAP	38	34	3	0	0	0
JCP	26	16	1	0	0	0
JPSP	44	88	3	2	5	1
Total	108	138	7	2	5	1
Proportion (%)	41.4	52.9	2.7	0.8	1.9	0.4

Note: The proportion reflects N =261 means tests. The sum of proportions did not exactly equal 100% because of rounding errors.

The use of ANCOVA is not as common as the other two types of means tests in the three reviewed journals. Thompson (2004) cautioned researchers about ANCOVA applications given the method's very strict assumptions (i.e., homogeneity of regression, extremely reliable measurement of covariates, and interpretable residualized dependent variables) that are very difficult to meet, and noting that the misuse of ANCOVA "may result in the analysis of an uninterpretable outcome variable" (p. 358).

Assessment of validity assumptions. When researchers use ANOVA as an analytic technique, a very important first step is to verify the distributional assumptions. If these assumptions are not reasonably met, results generated from the ANOVA test will "at best, [be] somewhat different from what they should be and, at worst, worthless" (Keselman et al., 1998, p. 351). However, among all 261 means tests, only 3 means tests addressed the violation of assumptions and used alternative tests to obtain more robust results. I was unable to determine whether or not most means tests assumptions were satisfied because most of the authors failed to address the issue. As shown in Table 6, most researchers (94.3%) neglected to provide *any* information on statistical assumptions tested when using means tests to evaluate group differences. Only 2.7% (7) of the authors addressed the possible violation of homogeneity assumptions, 4.6% (12) addressed the possible violation of normality, and no authors addressed the independence assumption (see Table 6).

Researchers seemed more concerned about non-normality issues, even though heterogeneity of variance within an unbalanced design may result in more severe departures from the true values (Glass et al., 1972; Skidmore & Thompson, 2013). When authors reported using tests that evaluated ANOVA assumptions, authors usually did not report the name of the tests that were used to assess the assumption. Only two means tests directly mention that Levene's test was used to test the homogeneity of variance assumptions, and these occurrences came from a single article. When assumptions were violated, the transformation was the most frequently reported resolution; eight tests used this adjustment to obtain a more robust estimate. Winsorizing and trimming were the

next most frequently used methods, which were reported four times in the total of 261counted means tests. One means test used nonparametric analysis, and two failed to report what procedure was used to address the violation issue.

Table 6

How Assumption Violations Were Addressed

What statistical violation assumptions were mentioned?			What tests (if any) were performed to test for / violation assumptions?			How were assumptions violations dealt with?		
	n	%		n	%		n	%
none	246	94.30%	None	246	94.30%	nothing was done because assumptions were not addressed	246	94.30%
independence of observations	0	0.00%	Levene's	2	0.80%	something was done but didn't mention what procedure was used	2	0.80%
homogeneity of variance	7	2.70%	Shapiro-Wilks	0	0.00%	transformation	8	3.10%
distribution (normality)	12	4.60%	Bartlett's test	0	0.00%	use of nonparametric analyses	1	0.40%
			Test was run but no name was given	13	5.00%	winsorizing and trimming	4	1.50%

Note. The percentage reflects the 261 means tests. The sum of percentage for the first column was not equal to 100% because one test addressed two types of violations.

The ways and levels for ANOVA means tests. Methodologists often conduct simulation studies to estimate how violation of assumptions will affect the accuracy of statistical results. However, before they think about the degree of deviation from

distributional assumptions (i.e., non-normality, heterogeneity of variance, and dependent residuals), another question to first address is "what is the most commonly used research design in the behavioral science research?" Because when simulation studies simulate the most commonly used designs, Monte Carlo simulation results can be of use to more applied researchers. Table 7 shows how the 261 documented means tests distributed based on the number of ways and levels. The most commonly used ANOVA was the one-way ANOVA, which comprised 80.8% (211) of the documented means tests. In the one-way ANOVA, 82.9% (175) were two-group mean difference tests, 11.4% (24) were three-group mean difference tests, 3.3% (7) were four-group mean difference tests, and 1.4% (3) did not provide enough information to be able to determine a description. The use of two-factor ANOVA means tests in psychological research was used quite often too, 17.2% (45) of documented means tests were two-way ANOVA means tests. Within the two-way ANOVAs, 86.7% were 2 × 2 ANOVA, 6.7% were 2 × 3 ANOVA, 2.2% were 2 × 4 ANOVA, and for 4.4%, it was impossible to determine. Three-way ANOVA and four-way ANOVA were only occasionally used in psychological research and usually had no more than two levels in each way.

ANOVAs appearing in articles were used in different ways. Some ANOVAs were used to answer the main research question (e.g., an ANOVA test for the ANOVA research design), while other ANOVAs were used for testing the preliminary condition (e.g., if the results differ by gender, or if the drop-off students were different from the examined participants). All 261 ANOVA means tests were further coded. The former

ANOVA tests were coded as “main research question ANOVA test”, and the latter ANOVA tests were coded as “manipulation check ANOVA test”.

Among one-way ANOVA means tests, 95 out of 211 were used for main research questions; while among two-way ANOVA means tests, 36 out of 45 were used for main research questions. And all three or more way ANOVA means tests were used for main research questions.

Table 7

Frequency of Ways and Levels for Reported ANOVAs

Number of Ways	Frequency	Percentage	Number of Levels	Frequency	Percentage
One way	211	80.8%	Two-group	175	82.9%
			Three-group	24	11.4%
			Four-group	7	3.3%
			Five-group	2	0.9%
			Not mentioned	3	1.4%
Two ways	45	17.2%	2 × 2	39	86.7%
			2 × 3	3	6.7%
			2 × 4	1	2.2%
			2 × ?	2	4.4%
Three ways	4	1.5%	2 × 2 × 2	4	100.0%
Four ways	1	0.4%	2 × 2 × 2 × 2	1	100.0%

Note. The percentage for the number of factors reflects the total number of documented means tests. And the percentage for the number of levels reflects the number of means tests that have the same number of ways.

Group size. Researchers did not pay much attention to group sizes when using ANOVAs to evaluate their research goals. More than 80% (210) of the tests provided no information about the group sizes when reporting ANOVA results. Among the 51 means tests where group sizes were discernible, only 6 means tests were balanced (participants were equal across groups), the other 45 means tests all had unequal group sizes. Of those unbalanced designs, 29 had a group size ratio (from high to low) smaller than 2, 18 had the ratios between 2 and 10, and 5 had the ratios larger than 10. The largest ratio in an unbalanced ANOVA observed in the reviewed articles was 70 (Boswell, McAleavey, Castonguay, Hayes, & Locke, 2012)! The distribution of group size ratio is showing in the Figure 2.

Variance. Compared to the group size, more researchers paid attention to the standard deviation or variance for each group. However, the overall number of means tests that reported information about standard deviation or variance was still small. Only 30% (80) of the means tests reported the standard deviation for each group. Among the 80 means tests that reported the standard deviation, 63 means tests had a ratio of the standard deviation (from high to low) smaller than 1.5, 12 means tests had a ratio between 1.5 to 2, and 5 means tests had a ratio greater than 2. The largest ratio of standard deviation in the coded articles was 8.2 (Rico, Sánchez-Manzanares, Antino, & Lau, 2012)! However, very few researchers conducted tests (e.g., Levene's, Shapiro-Wilk, or Bartlett's tests) to validate the homogeneity of variance assumption. The authors might not realize that when the standard deviation of each group varies too

much, the *p*-values and effect sizes (e.g., $r^2$and η^2) generated are essentially meaningless. The distribution of standard deviation ratio is showing in the Figure 2.

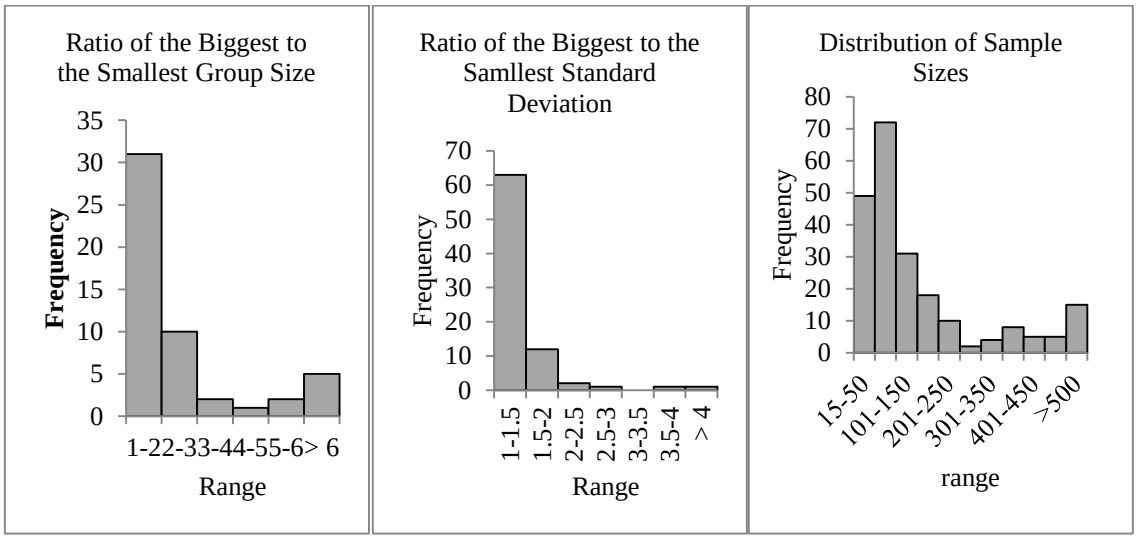

Figure 2. The distributions of ratios for group size and standard deviation (from the largest to the smallest) and the distribution of sample sizes.

Sample sizes. The given sample size reported in a study is often different that the sample size reported in a given analysis as variables used in each analysis can have different levels of missingness. Therefore, as *F* tests are the object of interest, I will discuss the sample size provided by the *F* test. Most coded ANOVA *F* tests had moderately large sample sizes. Among the 261 means tests, 220 reported the total sample sizes that ranged from 15 to 27,565, of which, 49 means tests had sample sizes that ranged between 15 and 50, 72 means tests had sample sizes that ranged between 51

to 100, 31 means tests had sample sizes that ranged between 101 to 150, 18 means tests had sample sizes that ranged between 151 to 200, and the other 49 means tests had sample sizes greater than 200. Only 8 means tests used sample sizes smaller than 30. The distribution of sample size is also showing in the Figure 2.

Pairing. "Paring" refers to the situation when heterogeneity of variance exists together with unequal group sizes. There are two types of "pairing" possible when using ANOVA to test the mean differences across unequal size groups: "positive pairing" is defined as the larger group having the larger variance and the smaller group having the smaller variance, while "negative pairing" is defined as larger group having smaller variance and smaller group having larger variance. Previous simulation studies have revealed that when negative pairings exist, estimates of effect sizes have positive sampling errors bias; when positive pairing exist, estimates of effect sizes have negative sampling errors bias (Skidmore & Thompson, 2013). However, very few ANOVAs provided enough information to determine which type of pairing categorized the data. Among the 261 tests, 18 ANOVAs provided enough information to discern 9 positive pairings and 9 negative pairings. For all other means tests I was unable to discern, because only the variance for each group was reported or only reported the sizes for each group, or neither group variance nor size was provided.

***p*-Value and Effect size.** Researchers' reliance on *p*-values when reporting the statistical results continues to dominate practice. Researchers often neglected to report the group sizes, the group variances, the validity of ANOVA assumptions, the effect sizes, and all other necessary components, but they never forgot to report the *p*-value,

regardless of whether they reported the value directly or reported it as a comparison to the critical value. All of the documented 261 means tests reported the *p*-values for ANOVA tests. Because multi-way ANOVAs may have more than one *F* tests with more than one corresponding *p*-values, I documented a total of 292 *p* values. of which, 36 *p* values were reported as "$p < .05$", 52 *p* values were reported as "$p < .01$", 50 *p* values reported as "$p < .001$", 13 *p* values were reported as "ns" without providing a value or range, and 11 *p*-values were reported as *p* greater than or smaller than a value other than the commonly used benchmarks (e.g., .05, .01, and .001). Only 95 reported the exact *p*-values. Dichotomous thinking about *p*-values (i.e., researchers mistakenly believe that whether study results were important or not was determined by whether $p_{calculate}$ is greater than $\alpha_{critical}$ or not (Thompson, 1989)) is still common in the investigated journals. In the worst cases, *p* values were only reported as "*ns*" (i.e., non-significant) or "significant" without providing the critical α as a criterion. Reporting *p*-value as "*ns*" or "significant" is essentially useless because for the same value $p = 0.03$, I can claim it as "*ns*", if I use the $\alpha_{critical} = 0.01$; I can also claim it as "significant", if $\alpha_{critical} = 0.05$.

Compared to the reporting of *p*-values, researchers' emphasis on effect sizes is apparently inadequate. Among the 261 means tests, 119 means tests reported the effect sizes, which is 45.6% of the total documented means tests. Among those that reported effect size, 44.5% reported partial η^2, 32.8% reported Cohen's *d*, 21.0% reported η^2, and 1.7% reported ξ^2. Partial η^2 is the most frequently reported effect size.

Post hoc test. Most of the documented means tests did not report any post hoc tests. For those that reported post hoc tests, 2 were LSD, 1 was Bonferroni, 6 were Tukey, and 17 were impossible to determine.

Conclusions and Recommendations Concerning the ANOVA Practices

This review reveals that ANOVA *F* tests are still widely used in a variety of contexts. However, the proportion of data-analytic methods that use ANOVAs has decreased comparatively (e.g., Edgington, 1974; Skidmore & Thompson, 2010). Fixed-effects ANOVAs are the most popular inferential technique; however, random-effects and mixed-effects ANOVAs were also frequently used. One-way ANOVA *F* tests were the most frequently used inferential statistical methods, however, multi-way ANOVA *F* tests were more often used to address to the major research questions. But most Monte Carlo simulation studies have been based on one-way ANOVA practices (e.g., Glass et al., 1972; Skidmore & Thompson, 2013); very few Monte Carlo simulations studied multi-way ANOVA *F* tests.

The use of ANOVA has been around since the early 1920's (David, 1995) and the earliest recommendation from methodologists regarding the use of ANOVA probably can be traced back to the mid- twentieth century (Glass et al., 1972). But the practices of ANOVA *F* tests in three reviewed journals are still problematic. This review revealed that researchers in the behavioral sciences still rarely verify whether ANOVA assumptions are satisfied. They generally neglect to recognize that heterogeneity of variance and unequal sample sizes can seriously affect the accuracy of estimates of population parameters. For those researchers who attended to the importance of

verifying ANOVA assumptions, non-normality was more likely to be of concern than heterogeneity of variance, even though variance heterogeneity affects Type I errors more than non-normality does (see Table 6).

Reporting effect sizes along with *p*-values has remarkably increased. During the time Keselman et al. (1998) conducted their review (i.e., articles published in the 1994 or 1995 issues) effect sizes were almost never reported. But at the time I conducted the review (i.e., year 2012) nearly half ANOVA tests reported effect sizes along with *p*-values. But, no researchers reported confidence intervals, not to mention confidence intervals for effect sizes, even though many methodologists have recommended using confidence intervals to replace statistical significance tests (e.g., Meehl, 1997; Thompson, 1999, 2002). And for those who reported effect sizes, partial η^2 is the most commonly reported effect sizes. However, η^2 has been shown to have the largest positive sampling error bias when compared to ω^2 and ξ^2 effect sizes (Skidmore & Thompson, 2013). Clearly, there remains substantial room for improvement in contemporary analytic practice.

AN EXAMINATION ON ASSUMPTION VIOLATIONS AFFECT THE ESTIMATES OF PRACTICAL SIGNIFICANCE IN TWO-WAY FIXED-EFFECTS ANOVA

Although analysis of variance (ANOVA) has been around since the early twentieth century (David, 1995), it remains a popular inferential analysis for both between-subjects univariate designs and psychometric variance partitioning (Kane, 2002; Keselman et al., 1998). It is well known that the validity of traditional *F* tests of mean differences in ANOVA is based on three core assumptions: independence of observations, normally distributed populations, and homoscedasticity. If these assumptions are violated, "it can be—and has been—shown that the resulting significance probabilities (*p*-values) are, at best, somewhat different from what they should be and, at worst, worthless" (Keselman et al., 1998, p. 351). Strictly speaking, it is appropriate to apply ANOVA only when three core assumptions are met. However, in practice, data never perfectly meet these three assumptions, and "the question is not whether ANOVA assumptions are perfectly met but, rather, whether assumptions are sufficiently well met that reasonable confidence can be vested in the ANOVA statistics" (Skidmore & Thompson, 2013, p. 536). Therefore, Monte Carlo simulations have been frequently used to estimate the extent of impact of assumptions violations on the validity of ANOVA tests.

Some Monte Carlo simulation studies addressed the consequences of nonindependence that affected the validity of ANOVA tests. For example, Harwell (1991) noted that "unequal correlations among errors can produce significance tests with inflated or conservative Type I error rates or tests with poor power" (p. 84). Yu (1995)

reported that correlated errors deflated the actual Type I error rate in the null condition, but the actual power remained unchanged or even higher than theoretical levels when the theoretical power increased to a certain level in the non-null condition. And Hurst (1996) found a high risk of the presence of violations to the assumption of independence in research on couples.

Studies on the violation of the normality and homogeneity of variance assumptions are also well documented. Harwell et al. (1992) meta-analyzed 28 Monte Carlo studies for the one- and two- factor, fixed-effects ANOVA model from three major databases, and summarized the effects of the assumption violations on Type I error rates and on the power of the *F* test. The findings supported the conclusion that the *F* test is relatively robust to "mild departure[s] from normality" but slightly affected by "moderately non-normal distribution[s]" (p. 316). The effect of the violation of the assumption of equal variance is confounded with sample sizes: negative pairings (e.g., small sample sizes paired with large variance) produce an inflated α rate, and positive pairings (e.g., small samples paired with small variances) produce conservative α rates, but equal sample sizes "mitigate" (p. 317) the influence of unequal variance on α .

The existing simulation studies overwhelmingly focused on violations of the normality and homogeneity of variance assumptions and their influence on power and *p*-values in null hypothesis statistical significance testing (NHSST), but neglect "a second and at least equally important use of the ANOVA" (Skidmore & Thompson, 2013, p. 537)—the estimate of practical significance, i.e., effect size. Such an omission implicitly asserts that practical significance is not as important as NHSST. Even though statistical

significance testing "provides information on the likelihood of finding the observed relationship by chance alone (sampling error)" (Olejnik & Algina, 2000, p. 241), *p* is a confounded index of sample size, true population differences, power, etc. Therefore, a single *p* value tells you nothing about how big the true difference is. Conversely, effect sizes directly estimate the magnitude of the mean differences. Effect sizes help researchers interpret statistically "significant" results with trivially observed small differences but huge sample sizes, and statistical "nonsignificant" results with moderate observed differences but very small sample sizes. Psychologists have increasingly emphasized the importance of effect sizes. The *Publication Manual of the American Psychological Association*, 6th edition (American Psychological Association, 2010) pointed out that NHSST is only a "starting point" (p. 33), and effect sizes "are needed to convey the most complete meaning of results" (p. 33). The lack of empirical work on effect sizes may have gone unnoticed in previous years, but with the increasing emphasis on effect sizes we can no longer continue to ignore the need for research in this area. Skidmore and Thompson (2013) helped fill the gap with an empirical study of one-way multiple group designs (k = 2, 3, and 4). Skidmore and Thompson (2013) demonstrated that when heterogeneity of variance in unbalanced designs was combined with negative pairing, $\hat{\eta}^2$, $\hat{\varepsilon}^2$, and $\hat{\omega}^2$ all tended to have positive sampling error biases, and when heterogeneity of variance in unbalanced designs combined with positive pairing, $\hat{\varepsilon}^2$, and $\hat{\omega}^2$ tended to have negative sampling error bias.

Until now, very few simulation studies have explored the effect of violation of ANOVA assumptions on practical significance in the multi-factor case. Not only are

multi-factor ANOVA simulation studies focusing on the effect of a violation of assumptions on effect size extremely rare, simulation studies of multi-factor ANOVA focusing on the power and *p*-values for NHSST are also uncommon. Indeed, in a review of Monte Carlo ANOVA simulation studies, Harwell, Rubinstein, Hayes, and Olds (1992) located four times as many one-way designs as two-factor models. Among the few papers exploring two-way designs, most were focused on nonorthogonal analysis of variance (Luh & Guo, 2001; Milligan, Wong, & Thompson, 1985, 1987). No two-way ANOVA simulations, to date, have focused on the impact of assumption violations on estimates of practical significance. Given that the conclusions derived from the single-factor case may or may not generalize to the multiple factors case, some unique elements in multi-way ANOVA, such as interaction factors, and additional simulation conditions (e.g., number of factors and number of levels in each factor), still deserve investigation.

The purpose of the present study was to examine the robustness of estimates of practical significance (i.e., $\hat{\eta}^2$, partial $\hat{\eta}^2$, $\hat{\varepsilon}^2$, partial $\hat{\varepsilon}^2$, $\hat{\omega}^2$, and partial $\hat{\omega}^2$) in a 2 × 3 two-way fixed-effects ANOVA. The study extended the exploration of these effect sizes in the presence of assumption violations, and is intended to generalize to the more common case of multi-factor ANOVAs.

Method

The conditions chosen for a good simulation study need to mirror actual research practices so that the results generated in simulations can be reasonably utilized to guide current research efforts. Therefore, conditions for the present simulation study

referenced both previous similar simulation studies and the findings from my review of ANOVA practices in three APA journals and other review studies.

Estimates of Practical Significance

The NHST in ANOVA *F* tests provides no direct information about the estimated population mean differences for main and interaction effects. There are three types of effect sizes: eta squared (η^2), epsilon squared (ε^2), and omega squared (ω^2) that are frequently used to measure practical significance in ANOVA. Eta squared is defined as the proportion of the total population variance that is accounted for by the variation in the dependent variable of interest (Grissom & Kim, 2012; Thompson, 2006). There are two formulas for η^2 to be used in fixed-effects ANOVA: (1) the classical formula proposed by Kerlinger (1973, p. 230), $\hat{\eta}^2 = \frac{SS_{effect}}{SS_{total}}$, and (2) the alternative formula proposed by Cohen, $\hat{\eta}_p^2 = \frac{df_{effect} F_{efect}}{df_{effect} F_{effect} + df_{error}}$ (Cohen, 1965). The total proportion of variance effect sizes and the partial proportion of variance effect sizes are not always equivalent: the former "assessed in terms of their contribution to the total variance,… are additive to the total proportion of 'explained' variance" while the latter is a more appropriate "measure of effect size for the factors under study" (Cohen, 1973, p. 109). In the one-way design, in the absence of other factors, the two formulas are equivalent; but when multiple factors exist, the denominator in the classical formula includes all factors, factor interactions and error variances, while the denominator in the alternative formula partitions out all other factors and interaction variances that are not of interest.

Eta squared or partial eta square (η_p^2) can be calculated directly using statistical software, like SPSS and SAS. Thus they are frequently reported as the measure of ANOVA effect sizes. However, many researchers incorrectly report both types of effect sizes. Further, the mislabeling of η^2 as η_p^2 in early versions of SPSS further exacerbated the confusion between the two (Levine & Hullett, 2002; Pierce, Block, & Aguinis, 2004).

It is well known that $\hat{\eta}^2$ is positively biased and tends to overestimate the parameter η^2 because sampling errors inflate the estimated SS_{effect}. There are two alternatives that researchers can choose to report: epsilon squared (ε^2) and omega squared (ω^2), which adjust for inflation due to sampling error. The equations are $\hat{\varepsilon}^2 = \frac{df_{effect}\left(MS_{effect} - MS_{error}\right)}{SS_{total}}$ and $\hat{\omega}^2 = \frac{df_{effect}\left(MS_{effect} - MS_{error}\right)}{SS_{total} + MS_{error}}$. Based on a Monte Carlo investigation (Keselman, 1975), the ε^2 is "a more nearly unbiased estimator" and the "bias of ω^2 is minimal" (Grissom & Kim, 2012, p. 182). The formula for ε^2 corrects the numerator of η^2 by subtracting the mean square error from the mean square effect, and ω^2 further adjusts the denominator of η^2 by adding the mean square error to the total sum of squares.

Epsilon squared and ω^2 also have alternative forms that partition out "all other nonerror sources of variances (main effects, interactions, trend components, etc.)" (Cohen, 1973, p. 108): partial epsilon squared (ε_p^2), and partial omega squared (ω_p^2).

The formula for ε_p^2 has the same denominator as η^2 and the same numerator as ε^2 (i.e., $\hat{\varepsilon}_p^{\,2} = \frac{df_{effect}\left(MS_{effect} - MS_{error}\right)}{SS_{effect} + SS_{error}}$), while the formula for ω_p^2 contains an adjustment of the degrees of freedom for error in the denominator $\hat{\omega}_p^{\,2} = \frac{df_{effect}\left(MS_{effect} - MS_{error}\right)}{SS_{effect} + \left(N - df_{effect}\right)MS_{error}}$ (Olejnik & Algina, 2000, p. 268).

The present study examined all proportion of variance effect sizes for ANOVAs (i.e., η^2, ε^2, ω^2, η_p^2, ε_p^2, and ω_p^2). The present investigation considered bias in effect sizes under violations of homoscedasticity and in combination with an unbalanced sampling strategy (i.e., positive pairing and negative pairing). Also, the classical forms (i.e., η^2, ε^2, and ω^2) and alternative forms (i.e., η_p^2, ε_p^2, and ω_p^2) were investigated to determine the influence of the violation conditions.

Population Effect Sizes Used in the Simulation

The *F* ratio, the test statistic for ANOVA, allows for testing the equality of multiple means for fixed effects under a variety of conditions (e.g., one-way ANOVA or multiple-way factorial designs under balanced or unbalanced conditions). The standardized overall effect size Cohen's *f* is commonly used to index the degree of departure from no effect. The value of *f* is given by $f = \frac{\sigma_\mu}{\sigma}$, wherein σ_μ is "the standard deviation of all of the means of the populations that are represented by the samples, and σ is the common standard deviation within each population" (Grissom, 2012, p. 180).

Cohen's *f* can be understood either as the "standard deviation of the standardized *k* population means" (Cohen, 1988, p. 276) or the correlation ratio. In one-way ANOVA or tests of main effects in factorial and other complex designs, Cohen's *f* can be translated to and from the Cohen's *d*, "the range of the standardized means, i.e., the distance between the smallest and largest of the *k* means"(Cohen, 1988, p. 276). However, such translation is not valid if the test of interactions is included in a factorial design. Therefore, in the present two-way factorial simulation study with both main and interaction effect, Cohen's *f* is used alone to quantify the effect size and is defined as "the ratio of the variance of the means to the variance of the values within the population" (Cohen, 1988, p. 281). The specific formula to describe the relationship between the *f* and the ratio of variance is $f^2 = \frac{\sigma_m^{\ 2}}{\sigma^2}$.

Cohen (1988) defined the benchmarks for "small", "medium", and "large" *f* values, respectively, $f = 0.10$, $f = 0.25$, and $f = 0.40$. Even though he emphasized at the end of his book, that "the values chosen had no more reliable a basis than my own intuition" (Cohen, 1988, p. 532) and researchers should "not employ them if possible" (Cohen, 1988, p. 532), the benchmark is arguably suitable in use of power analysis, meta-analysis, and simulation studies of population parameter settings. Because "values of *f* as large as 0.50 are not common in behavioral science" (Cohen, 1988, p. 284), the non-null conditions used in the present study are $f = 0.10$, $f = 0.25$, and $f = 0.40$. Including the null condition, four different Cohen's *f*s were used in this simulation study.

Group Means Used for Different Cohen's f Values

In a 2 × 3 ANOVA design, the Cohen's *fs* for A-way, B-way, and their interaction are determined by the following twelve components: the six cells' variances (i.e., S_{11}^2, S_{12}^2, S_{13}^2, S_{21}^2, S_{22}^2, and S_{23}^2) and the six cells' means (i.e., M_{11}, M_{12}, M_{13}, M_{21}, M_{22}, and M_{23}).

		B-way			
		B1	B2	B3	
A-way	A1	S_{11}^2	S_{12}^2	S_{13}^2	$S_{1.}^2$
		M_{11}	M_{12}	M_{13}	$M_{1.}$
		n_{11}	n_{12}	n_{13}	$N_{1.}$
	A2	S_{21}^2	S_{22}^2	S_{23}^2	$S_{2.}^2$
		M_{21}	M_{22}	M_{23}	$M_{2.}$
		n_{21}	n_{22}	n_{23}	$N_{2.}$
		$S_{.1}^2$	$S_{.2}^2$	$S_{.3}^2$	$S_{..}^2$
		$M_{.1}$	$M_{.2}$	$M_{.3}$	$M_{..}$
		$N_{.1}$	$N_{.2}$	$N_{.3}$	N

Unlike the one-way ANOVA that allows the *k* group means to be of any value, in a two-way factorial design the main and interaction effect means need to satisfy the following two constraints: (1) $\frac{\sum M_{i.}N_{i.}}{\sum N_{i.}} = \frac{\sum M_{.j}N_{.j}}{\sum N_{.j}} = M_{..}$, wherein $M_{i.}$ and $M_{.j}$ are the main effect means, $N_{i.}$ and $N_{.j}$ are each main effect's group sizes, and $M_{..}$ is the grand mean. For a balanced factorial design, this formula could be simplified as $\frac{\sum M_{i.}}{r} = \frac{\sum M_{.j}}{c} = M_{..}$, r and c represent the number of levels for the Row and Column

effect; and (2) $X_{ij} = M_{ij} - M_{i.} - M_{.j} + M_{..}$, wherein M_{ij} is the cell mean, and X_{ij} is the interaction effect mean (Cohen, 1988; Thompson, 2006). When Cohen's $f = 0$ for all factors and the interaction, the $M_{11} = M_{12} = M_{13} = M_{21} = M_{22} = M_{23} = M_{..} = 0$ (to simplify the computation, I constrained the grand mean equal to 0, but in the general case, a grand mean equal to zero is not a necessary condition). The $S_{.1}^2$ is the pooled variance of S_{11}^2, S_{12}^2, and S_{13}^2 ($S_{.1}^2 = \frac{S_{11}^2 + S_{12}^2 + S_{13}^2}{3}$), $S_{.2}^2$ is the pooled variance of S_{21}^2, S_{22}^2, and S_{23}^2, $S_{1.}^2$ is the pooled variance of $S_{11}^2$and S_{21}^2, $S_{2.}^2$ is the pooled variance of $S_{12}^2$and S_{22}^2, and $S_{3.}^2$ is the pooled variance of $S_{13}^2$and S_{23}^2. When Cohen's $f > 0$, the six cells' means no longer all equal 0, the total σ_{A1}^2 was then composed of two components, the within variance in each cell ($S_{.1}^2 = \frac{S_{11}^2 + S_{12}^2 + S_{13}^2}{3}$), and the between variance for the three cells $\frac{(M_{11} - M_{1.})^2 n_{11} + (M_{12} - M_{1.})^2 n_{12} + (M_{13} - M_{1.})^2 n_{13}}{n_{11} + n_{12} + n_{12} - 1}$. Because the between variance inflates the total σ_{A1}^2, $f^2 = \frac{\sigma_m^2}{\sigma^2}$ for the A-way is not only influenced by the A-way means variance. Any change on the B-way and interaction causes a change in the total variance, σ^2, and further changes Cohen's f for the A-way. Therefore, the 12 components (i.e., the six cells' variances and the six cells' means) need to be determined simultaneously.

Table 8 provides the standard deviations and the means for the six cells. To simplify the investigation, for each condition, the A-way, B-way and AB interaction all were assigned the same Cohen's f effect sizes. The grand mean was constrained to be 0,

whereby any change of main effect would not influence the other main and interaction effects; meanwhile, the interaction effect was constrained by the sum of columns equal to zero and the sum of rows equal to zero, whereby any change of interaction would not affect all main effect parameters.

Table 8

Cell Means and Cell Standard Deviations Used for Different Cohen's f

Cohen's *f*	S_{11}	S_{12}	S_{13}	S_{21}	S_{22}	S_{23}	M_{11}	M_{12}	M_{13}	M_{21}	M_{22}	M_{23}
0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.000	0.000	0.000	0.000	0.000	0.000
0.00	1.00	1.00	1.50	1.50	1.50	2.25	0.000	0.000	0.000	0.000	0.000	0.000
0.00	1.00	1.00	2.00	2.00	2.00	4.00	0.000	0.000	0.000	0.000	0.000	0.000
0.10	1.00	1.00	1.00	1.00	1.00	1.00	-0.345	-0.100	0.145	0.100	0.100	0.100
0.10	1.00	1.00	1.50	1.50	1.50	2.25	-0.523	-0.152	0.220	0.152	0.152	0.152
0.10	1.00	1.00	2.00	2.00	2.00	4.00	-0.771	-0.224	0.324	0.224	0.224	0.224
0.25	1.00	1.00	1.00	1.00	1.00	1.00	-0.862	-0.250	0.362	0.250	0.250	0.250
0.25	1.00	1.00	1.50	1.50	1.50	2.25	-1.308	-0.379	0.550	0.379	0.379	0.379
0.25	1.00	1.00	2.00	2.00	2.00	4.00	-1.928	-0.559	0.810	0.559	0.559	0.559
0.40	1.00	1.00	1.00	1.00	1.00	1.00	-1.380	-0.400	0.580	0.400	0.400	0.400
0.40	1.00	1.00	1.50	1.50	1.50	2.25	-2.094	-0.607	0.880	0.607	0.607	0.607
0.40	1.00	1.00	2.00	2.00	2.00	4.00	-3.085	-0.894	1.296	0.894	0.894	0.894

Variance and the Variance Ratios

Heterogeneity of variance has been reported to distort the parameter estimates in many simulation studies (Harwell, Rubinstein, Hayes, & Olds, 1992; Lix, Keselman, & Keselman, 1996). Therefore, the degree of departure from variance homogeneity is an important condition in my simulation study. In my ANOVA review study, I found in the

ANOVA studies that reported the standard deviation for each group (80 out of 261 means tests), 79% had a mean ratio from the largest to the smallest standard deviation equal or smaller than 1.5, 15% had a mean ratio between 1.5 to 2, very few had ratios greater than 2, and the median ratio was 1.4. Previous review studies reported similar findings. For example, Keselman et al. (1998) examined 85 factorial studies "with a mean ratio of 2.8 (*SD* = 4.2), a median of 1.7, and a maximum ratio of 29.4" (p. 356). Therefore, the three variance conditions chosen in the present study were: fully homogenous (variance ratio: A-way 1:1; and B-way 1:1:1); mild departure from homogeneity of variance (variance ratio: A-way 1: 2.25, B-way 1:1:2.25); and moderate departure of homogeneity of variance (variance ratio: A-way 1:4, B-way 1:1:4). It is worth noting that the variance ratio refers to the error variance ratio that has excluded systematic differences for the main effects and interactions. If I included the variation due to main effects, the actual variance ratio for each way is lower than the condition value because both the numerator and denominator add to the systematic between group variance and further deflate the variance ratio. Table 8 also provides the standard deviations for each cell. For the various effect size (Cohen's *f*) conditions, there were three types of variance ratios that defined the extent of the violation of the assumption of homogeneity that affected the validity of results.

Sample Sizes

Early simulation studies usually used relatively small sample sizes (average group sizes ranged from 1 to 9 with total sample sizes under 15) and the ratio from the largest to the smallest ranged from 1 to 5 (Glass, Peckham, & Sanders, 1972). But in my

review of ANOVA practices, I found no ANOVA *F* tests had sampled sizes smaller than 15; the sample sizes ranged from 15 to 27,565, with the median = 90, and more than 60% had size ratios smaller than 2. Therefore, I chose 6 and 36 as the average cell sizes (the corresponding total sample size is 36 and 216) and simulated both balanced and unbalanced cases. In the unbalanced situation, the group size ratios were 1:2 for the A-way, and 1:1:2 for the B-way.

Pairings

Previous simulation studies discovered that when unequal group sizes and heterogeneity of variance existed concurrently, the pairing (i.e., large sample sizes paired with large variances or large sample sizes paired with small variances) produced different results. Therefore, the present study considered both positive pairing (large sample sizes with large variance), and negative pairing (large sample sizes with small variance).

Replications

For a two-way Monte Carlo simulation study, a typical number of replications found in the sample programs in the SAS Monte Carlo studies guide book was 5,000 to 10,000 replications (Fan, Sivo, & Keenan, 2002). I tested the Type I error rate in the null condition with all assumptions satisfied using 5,000 replications. The actual type I error rate obtained from the simulation study ranged from 0.046 to 0.055. To ensure more stable results, I choose to run 10,000 replications for each condition. This analysis confirmed correct coding of the simulation syntax.

Table 9

Empirical Type I Error Rate and Empirical Experimentwise Error Rate for Normally Distributed Samples

Pairing	SD Ratio	Average Cell Sizes	Empirical Type I Error Rate			Experimentwise Error Rate
			A-way	B-way	Interaction	
Balanced	1:1(1:1:1)	6	0.052	0.052	0.048	0.141
Balanced	1:1(1:1:1)	36	0.052	0.051	0.050	0.145
Balanced	1:1.5 (1:1:1.5)	6	0.054	0.053	0.060	0.146
Balanced	1:1.5 (1:1:1.5)	36	0.049	0.056	0.051	0.141
Balanced	1:2 (1:1:2)	6	0.057	0.068	0.067	0.152
Balanced	1:2 (1:1:2)	36	0.050	0.064	0.064	0.146
Negative Pairing	1:1 (1:1:1)	6	0.055	0.049	0.051	0.139
Negative Pairing	1:1 (1:1:1)	36	0.046	0.051	0.045	0.131
Negative Pairing	1:1.5 (1:1:1.5)	6	0.114	0.132	0.137	0.284
Negative Pairing	1:1.5 (1:1:1.5)	36	0.108	0.128	0.127	0.273
Negative Pairing	1:2 (1:1:2)	6	0.179	0.225	0.217	0.386
Negative Pairing	1:2 (1:1:2)	36	0.158	0.189	0.193	0.352
Positive Pairing	1:1 (1:1:1)	6	0.046	0.048	0.053	0.131
Positive Pairing	1:1 (1:1:1)	36	0.050	0.050	0.047	0.136
Positive Pairing	1:1.5 (1:1:1.5)	6	0.010	0.011	0.010	0.030
Positive Pairing	1:1.5 (1:1:1.5)	36	0.011	0.011	0.011	0.032
Positive Pairing	1:2 (1:1:2)	6	0.004	0.007	0.004	0.015
Positive Pairing	1:2 (1:1:2)	36	0.003	0.005	0.004	0.011

Note. Numbers in parentheses reflect the standard deviation ratios for the three levels in B-way. The three standard deviation ratios for the six cells (SD_{C11}: SD_{C12}:SD_{C13}: SD_{C21}: SD_{C22}: SD_{C23}) were 1: 1: 1: 1: 1: 1, 1: 1: 1.5: 1.5: 1.5: 2.25, 1: 1: 2: 2: 2: 4, respectively.

In all, the simulation study considered four Cohen's *f* values (0, 0.10, 0.25, and 0.40), three variance ratios (1:1 for the A-way and 1:1:1 for the B-way; 1:1.5 for the A-way and 1:1:1.5 for the B-way; and 1:2 for the A-way and 1:1:2 for the B-way), three types of average cell sizes (6 and 36), three types of pairings (balanced, positive pairing, and negative pairing). Thus, there were $4 \times 3 \times 2 \times 3 = 72$ conditions with $72 \times 10{,}000 = 720{,}000$ total replications.

Simulation Baseline Check

Glass, Peckham, and Sanders (1972) emphasized the importance of baseline checks for conducting a robust simulation study. A series of "baseline check[s]" tests were carried out under all investigated conditions. When all assumptions are satisfied, the "actual and theoretical probability should be equal within sampling error" (p. 282).

In my baseline check Monte Carlo simulations, I investigated the empirical Type I error rate, the empirical power, and the empirical experimentwise error rate. Table 9 provides empirical Type I error rates under tested conditions. The nominal α was set to .05. When the homogeneity of variance assumption was perfectly satisfied, regardless of whether or not the group sizes in each way were equal, all empirical Type I error rates were close to 0.05. As long as the designs were balanced, the heterogeneity of variance did not have much influence on empirical Type I error rate. But when heterogeneity of variance existed together with the unbalanced group sizes, positive pairing deflated the empirical Type I error rate and negative pairing inflated the empirical Type I error rate. The results were consistent with previous simulation studies. Bathke (2004) reported that the ANOVA *F* test is robust in balanced designs with unequal variances and non-normal

data. Hsu (1938), Box (1954), and Horsnell (1953) all reported "that negatively pairing unequal sample sizes and variances… produces an inflated α rate; positive pairings… produce conservative α rates" (Harwell et al., 1992, p. 317).

The theoretical experimentwise error rate can be calculated using the Bonferroni formula $\alpha_{Experimentwise} = 1-(\alpha_{Testwise})^k$ (Thompson, 2006). In a balanced two-way factorial ANOVA, when nominal α was predetermined as 0.05, the theoretical experimentwise error rate $\alpha_{Experimentwise} = 1-(1-0.05)^3 = 0.143$. And the empirical experimentwise error rates obtained in my "baseline check" simulation were very close to the theoretical value. The empirical experimentwise error rates were also influenced by the unequal sample sizes and heterogeneous variance. When small samples paired with large variances, the empirical experimentwise error rates were strongly inflated, while when small samples paired with small variances, the empirical experimentwise error rates were strongly deflated.

Table 10 provides the empirical power estimates when assumptions were fully satisfied or with varying degrees of violations. As long as sample sizes were equal, the estimated empirical power is very close to the theoretical power obtained from the software G*Power 3.15, regardless of whether or not homogeneity of variance was satisfied. When the sample sizes were not equal across groups, the empirical power no longer remained constant. However, the estimated empirical power did not change monotonically by the types of pairing, and instead, was a confounded result of unequal variances, unequal sample sizes, and effect sizes. The results also matched the Harwell et

al. (1992) review of previous simulation studies, in which Harwell et al. concluded that "it is difficult to characterize general conclusions about power" (p. 317).

Results

The present simulation study obtained 720,000 random samples under 72 conditions. For each sample, the estimated $\hat{\eta}^2$, $\hat{\varepsilon}^2$, $\hat{\omega}^2$, $\hat{\eta}_p^2$, $\hat{\varepsilon}_p^2$, and $\hat{\omega}_p^2$ was computed. The estimated bias due to sampling error was calculated as the distance between each estimated effect size to the true population effect sizes. A positive bias indicates an overestimated effect size, and a negative bias indicates an underestimated effect size.

Parameter Bias

The bias is the difference between the estimated effect sizes and the true parameter value. For all of the 720,000 sample units, the estimated $\hat{\eta}^2$, $\hat{\varepsilon}^2$, $\hat{\omega}^2$, $\hat{\eta}_p^2$, $\hat{\varepsilon}_p^2$, and $\hat{\omega}_p^2$ were calculated and compared to the true population effect sizes. Eighteen (6 × 3) four-way full factorial ANOVAs were conducted using the least squares estimation methods with the outcome variables being the bias of $\hat{\eta}^2$, the bias of $\hat{\varepsilon}^2$, the bias of $\hat{\omega}^2$, the bias of $\hat{\eta}_p^2$, the bias of $\hat{\varepsilon}_p^2$, and bias of $\hat{\omega}_p^2$, respectively (6), for the A-way, B-way, and the AB interaction (3).

Table 10

Empirical Power Estimates with Normal Distribution and Different Sample Sizes

Cohen's *f*	Pairing	SD Ratio	Average Cell sizes	A-way Power		B-way Power		Interaction Power	
				Empirical	Theoretical	Empirical	Theoretical	Empirical	Theoretical
0.1	Balanced	1:1	6	0.085	0.090	0.074	0.076	0.075	0.076
0.1	Balanced	1:1.5	6	0.086	/	0.080	/	0.079	/
0.1	Balanced	1:2	6	0.098	/	0.097	/	0.098	/
0.1	Negative Pairing	1:1	6	0.084	/	0.074	/	0.072	/
0.1	Negative Pairing	1:1.5	6	0.156	/	0.175	/	0.167	/
0.1	Negative Pairing	1:2	6	0.220	/	0.257	/	0.258	/
0.1	Positive Pairing	1:1	6	0.082	/	0.072	/	0.074	/
0.1	Positive Pairing	1:1.5	6	0.026	/	0.023	/	0.021	/
0.1	Positive Pairing	1:2	6	0.012	/	0.013	/	0.012	/
0.25	Balanced	1:1	6	0.310	0.308	0.241	0.231	0.225	0.231
0.25	Balanced	1:1.5	6	0.308	/	0.232	/	0.237	/
0.25	Balanced	1:2	6	0.314	/	0.240	/	0.236	/
0.25	Negative Pairing	1:1	6	0.256	/	0.211	/	0.208	/
0.25	Negative Pairing	1:1.5	6	0.353	/	0.332	/	0.339	/
0.25	Negative Pairing	1:2	6	0.413	/	0.420	/	0.421	/
0.25	Positive Pairing	1:1	6	0.255	/	0.215	/	0.220	/
0.25	Positive Pairing	1:1.5	6	0.136	/	0.100	/	0.099	/
0.25	Positive Pairing	1:2	6	0.088	/	0.059	/	0.057	/
0.4	Balanced	1:1	6	0.636	0.645	0.516	0.525	0.516	0.525
0.4	Balanced	1:1.5	6	0.650	/	0.511	/	0.505	/
0.4	Balanced	1:2	6	0.640	/	0.498	/	0.508	/
0.4	Negative Pairing	1:1	6	0.548	/	0.484	/	0.482	/

Table 10 Continued

Cohen's *f*	Pairing	SD Ratio	Average Cell sizes	A-way Power		B-way Power		Interaction Power	
				Empirical	Theoretical	Empirical	Theoretical	Empirical	Theoretical
0.4	Negative Pairing	1:1.5	6	0.623	/	0.601	/	0.598	/
0.4	Negative Pairing	1:2	6	0.675	/	0.649	/	0.660	/
0.4	Positive Pairing	1:1	6	0.550	/	0.476	/	0.482	/
0.4	Positive Pairing	1:1.5	6	0.414	/	0.309	/	0.311	/
0.4	Positive Pairing	1:2	6	0.327	/	0.216	/	0.215	/
0.1	Balanced	1:1	36	0.309	0.310	0.236	0.238	0.235	0.238
0.1	Balanced	1:1.5	36	0.319	/	0.234	/	0.243	/
0.1	Balanced	1:2	36	0.310	/	0.239	/	0.239	/
0.1	Negative Pairing	1:1	36	0.252	/	0.216	/	0.220	/
0.1	Negative Pairing	1:1.5	36	0.346	/	0.330	/	0.331	/
0.1	Negative Pairing	1:2	36	0.386	/	0.395	/	0.395	/
0.1	Positive Pairing	1:1	36	0.263	/	0.225	/	0.229	/
0.1	Positive Pairing	1:1.5	36	0.148	/	0.106	/	0.103	/
0.1	Positive Pairing	1:2	36	0.085	/	0.058	/	0.060	/
0.25	Balanced	1:1	36	0.955	0.955	0.912	0.915	0.919	0.915
0.25	Balanced	1:1.5	36	0.956	/	0.898	/	0.899	/
0.25	Balanced	1:2	36	0.952	/	0.884	/	0.895	/
0.25	Negative Pairing	1:1	36	0.909	/	0.892	/	0.889	/
0.25	Negative Pairing	1:1.5	36	0.922	/	0.905	/	0.904	/
0.25	Negative Pairing	1:2	36	0.925	/	0.912	/	0.912	/
0.25	Positive Pairing	1:1	36	0.903	/	0.884	/	0.890	/
0.25	Positive Pairing	1:1.5	36	0.878	/	0.793	/	0.796	/
0.25	Positive Pairing	1:2	36	0.854	/	0.725	/	0.726	/

Table 10 Continued

Cohen's *f*	Pairing	SD Ratio	Average Cell sizes	A-way Power		B-way Power		Interaction Power	
				Empirical	Theoretical	Empirical	Theoretical	Empirical	Theoretical
0.4	Balanced	1:1	36	1.000	1.000	1.000	1.000	1.000	1.000
0.4	Balanced	1:1.5	36	1.000	/	1.000	/	0.999	/
0.4	Balanced	1:2	36	1.000	/	0.999	/	1.000	/
0.4	Negative Pairing	1:1	36	1.000	/	1.000	/	0.999	/
0.4	Negative Pairing	1:1.5	36	1.000	/	0.999	/	0.999	/
0.4	Negative Pairing	1:2	36	0.999	/	0.999	/	0.999	/
0.4	Positive Pairing	1:1	36	0.999	/	1.000	/	1.000	/
0.4	Positive Pairing	1:1.5	36	1.000	/	0.998	/	0.999	/
0.4	Positive Pairing	1:2	36	1.000	/	0.998	/	0.998	/

Note. Theoretical power values obtained using G* Power 3.15.

Table 11

Estimated Parameter Bias for A-Way Effect Sizes of η^2, ε^2, ω^2 in the 2 × 3 ANOVA

A-way		η^2		Partial η^2		ε^2		Partial ε^2		ω^2		Partial ω^2	
Source	*df*	SS	η^2	SS	η^2	SS	η^2	SS	η^2	SS	η^2	SS	η^2
Cohen's *f*	3	34.22	0.02	14.60	0.00	17.15	0.01	8.99	0.00	20.09	0.01	34.51	0.01
SD Ratio	2	0.37	0.00	2.48	0.00	0.46	0.00	2.63	0.00	0.47	0.00	2.34	0.00
Group Size Ratio	2	53.96	0.03	119.62	0.04	63.97	0.03	125.34	0.04	62.44	0.03	107.39	0.04
Total *N*	1	81.97	0.04	180.72	0.06	0.03	0.00	5.71	0.00	0.00	0.00	0.17	0.00
Cohen's *f* * SD Ratio	6	0.01	0.00	0.23	0.00	0.01	0.00	0.24	0.00	0.01	0.00	0.26	0.00
Cohen's f * Group Size Ratio	6	15.29	0.01	43.74	0.01	16.49	0.01	45.39	0.01	16.38	0.01	41.33	0.02
SD Ratio *Group Size Ratio	4	25.24	0.01	64.92	0.02	30.46	0.02	68.11	0.02	29.79	0.02	58.40	0.02
Cohen's f * Total *N*	3	1.75	0.00	0.88	0.00	0.01	0.00	2.45	0.00	0.11	0.00	0.24	0.00
SD Ratio * Total *N*	2	0.16	0.00	1.10	0.00	0.22	0.00	1.19	0.00	0.22	0.00	1.01	0.00
Group Size Ratio * Total *N*	2	4.31	0.00	14.10	0.00	7.13	0.00	15.79	0.00	6.73	0.00	10.96	0.00
Cohen's f * SD Ratio * group Size Ratio	12	3.37	0.00	16.58	0.01	3.67	0.00	17.20	0.01	3.73	0.00	16.06	0.01
Cohen's *f* * SD Ratio * Total *N*	6	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.00	0.01	0.00	0.03	0.00
Cohen''s *f* * Group Size ratio * Total *N*	6	0.53	0.00	0.15	0.00	0.44	0.00	0.14	0.00	0.41	0.00	0.13	0.00
SD Ratio * Group Size Ratio * Total *N*	4	2.21	0.00	8.69	0.00	3.69	0.00	9.73	0.00	3.49	0.00	6.78	0.00
Cohen's f * SD Ratio * Group Size Ratio * Total *N*	12	0.37	0.00	0.08	0.00	0.33	0.00	0.06	0.00	0.31	0.00	0.07	0.00
Error	719999	1649.09	0.88	2704.26	0.85	1725.10	0.92	2871.04	0.90	1668.62	0.92	2371.47	0.89
Total		1872.84		3172.16		1869.16		3174.02		1812.79		2651.15	

Table 12

Estimated Parameter Bias for B-Way Effect Sizes of η^2, ε^2, ω^2 in the 2 × 3 ANOVA

B-way		η^2		partial η^2		ε^2		partial ε^2		ω^2		partial ω^2	
Source	*df*	SS	η^2	SS	η^2	SS	η^2	SS	η^2	SS	η^2	SS	η^2
Cohen's *f*	3	23.99	0.01	6.49	0.00	2.20	0.00	0.80	0.00	3.45	0.00	10.52	0.00
SD Ratio	2	0.48	0.00	2.88	0.00	0.71	0.00	3.21	0.00	0.73	0.00	3.07	0.00
Total *N*	1	321.56	0.11	542.85	0.12	0.02	0.00	4.17	0.00	0.01	0.00	0.27	0.00
Group Size Ratio	6	0.08	0.00	0.06	0.00	0.08	0.00	0.06	0.00	0.08	0.00	0.06	0.00
Cohen's *f* * SD Ratio	2	100.20	0.03	200.46	0.05	129.26	0.05	221.01	0.05	125.72	0.05	194.02	0.05
Cohen's f * Group Size Ratio	6	7.32	0.00	30.59	0.01	9.08	0.00	32.76	0.01	9.19	0.00	31.34	0.01
SD Ratio *Group Size Ratio	4	54.56	0.02	117.42	0.03	70.17	0.02	129.49	0.03	68.36	0.02	113.95	0.03
Cohen's f * Total *N*	12	2.84	0.00	15.08	0.00	3.42	0.00	16.11	0.00	3.50	0.00	15.61	0.00
SD Ratio * Total *N*	3	7.21	0.00	0.19	0.00	0.05	0.00	0.64	0.00	0.23	0.00	0.73	0.00
Group Size Ratio * Total *N*	2	0.13	0.00	1.14	0.00	0.25	0.00	1.34	0.00	0.26	0.00	1.27	0.00
Cohen's f * SD Ratio * group Size Ratio	6	0.05	0.00	0.07	0.00	0.05	0.00	0.08	0.00	0.05	0.00	0.05	0.00
Cohen's *f* * SD Ratio * Total *N*	2	14.85	0.01	35.30	0.01	25.41	0.01	43.15	0.01	24.04	0.01	33.17	0.01
Cohen's *f* * Group Size ratio * Total *N*	6	1.14	0.00	0.58	0.00	0.85	0.00	0.44	0.00	0.80	0.00	0.44	0.00
SD Ratio * Group Size Ratio * Total *N*	4	7.74	0.00	20.87	0.00	13.32	0.00	25.52	0.01	12.63	0.00	19.71	0.01
Cohen's f * SD Ratio * Group Size Ratio * Total *N*	12	0.96	0.00	0.46	0.00	0.82	0.00	0.36	0.00	0.77	0.00	0.36	0.00
Error	719999	2333.94	0.81	3437.63	0.78	2571.30	0.91	3870.22	0.89	2484.77	0.91	3314.21	0.89
Total		2877.05		4412.08		2826.99		4349.37		2734.57		3738.78	

Table 13

Estimated Parameter Bias for AB-Interaction Effect Sizes of η^2, ε^2, ω^2 in the 2 × 3 ANOVA

AB interaction		η^2		partial η^2		ε^2		partial ε^2		ω^2		partial ω^2	
Source	*df*	SS	η^2	SS	η^2	SS	η^2	SS	η^2	SS	η^2	SS	η^2
Cohen's *f*	3	14.37	0.00	5.77	0.00	0.15	0.00	0.55	0.00	0.62	0.00	9.64	0.00
SD Ratio	2	2.19	0.00	3.24	0.00	2.66	0.00	3.61	0.00	2.64	0.00	3.46	0.00
Total *N*	2	119.11	0.04	202.19	0.05	150.59	0.05	223.00	0.05	146.18	0.05	195.62	0.05
Group Size Ratio	1	343.80	0.11	542.08	0.12	0.58	0.00	4.10	0.00	0.26	0.00	0.26	0.00
Cohen's *f* * SD Ratio	6	0.50	0.00	0.06	0.00	0.55	0.00	0.06	0.00	0.53	0.00	0.08	0.00
Cohen's f * Group Size Ratio	6	12.44	0.00	30.95	0.01	14.61	0.00	33.17	0.01	14.59	0.00	31.69	0.01
SD Ratio *Group Size Ratio	4	63.79	0.02	121.46	0.03	80.60	0.03	134.02	0.03	78.45	0.03	117.87	0.03
Cohen's f * Total *N*	3	3.37	0.00	0.07	0.00	0.41	0.00	1.00	0.00	0.12	0.00	0.46	0.00
SD Ratio * Total *N*	2	0.69	0.00	1.13	0.00	0.93	0.00	1.34	0.00	0.92	0.00	1.26	0.00
Cohen's f * SD Ratio * group Size Ratio	12	5.25	0.00	16.57	0.00	6.02	0.00	17.76	0.00	6.08	0.00	17.12	0.00
Cohen's *f* * SD Ratio * Total *N*	6	0.20	0.00	0.05	0.00	0.23	0.00	0.06	0.00	0.22	0.00	0.04	0.00
Group Size Ratio * Total *N*	2	20.69	0.01	37.02	0.01	32.86	0.01	45.15	0.01	31.04	0.01	34.82	0.01
Cohen's *f* * Group Size ratio * Total *N*	6	0.17	0.00	0.38	0.00	0.06	0.00	0.24	0.00	0.05	0.00	0.27	0.00
SD Ratio * Group Size Ratio * Total *N*	4	10.40	0.00	22.64	0.01	16.75	0.01	27.61	0.01	15.89	0.01	21.41	0.01
Cohen's f * SD Ratio * Group Size Ratio * Total *N*	12	0.29	0.00	0.25	0.00	0.22	0.00	0.18	0.00	0.20	0.00	0.18	0.00
Error	719999	2641.42	0.82	3417.07	078	2838.65	0.90	3847.04	0.89	2742.24	0.90	3304.60	0.88
Total		3238.68		4400.93		3145.87		4338.88		3040.05		3738.78	

Tables 11, 12, and 13 provide the four-way full factorial ANOVAs results for A-way, B-way, and AB interaction. In general, the classical forms (i.e., $\hat{\eta}^2$, $\hat{\varepsilon}^2$, and $\hat{\omega}^2$) had smaller total sum of squares than the alternative forms (i.e., $\hat{\eta}_p^2$, $\hat{\varepsilon}_p^2$, and $\hat{\omega}_p^2$). Figure 3 provides the boxplots for the bias of $\hat{\eta}^2$ and the bias of $\hat{\eta}_p^2$ across heterogeneity, sampling type, and values of Cohen's *f* (the boxplots for the bias of $\hat{\varepsilon}^2$ and $\hat{\varepsilon}_p^2$, the bias of $\hat{\omega}^2$ and $\hat{\omega}_p^2$ were very similar to those displayed in Figure 3, and thus are omitted here). As long as the designs were balanced and had homogeneous variance, the $\hat{\eta}^2$ and $\hat{\eta}_p^2$ had similar mean parameter estimates, but the biases for the $\hat{\eta}_p^2$ were more "spread out" and had larger sums of squares. When an unbalanced design was paired with a heterogeneous variance, the $\hat{\eta}_p^2$ was more greatly affected by the unmet assumptions and tended to yield more biased estimates.

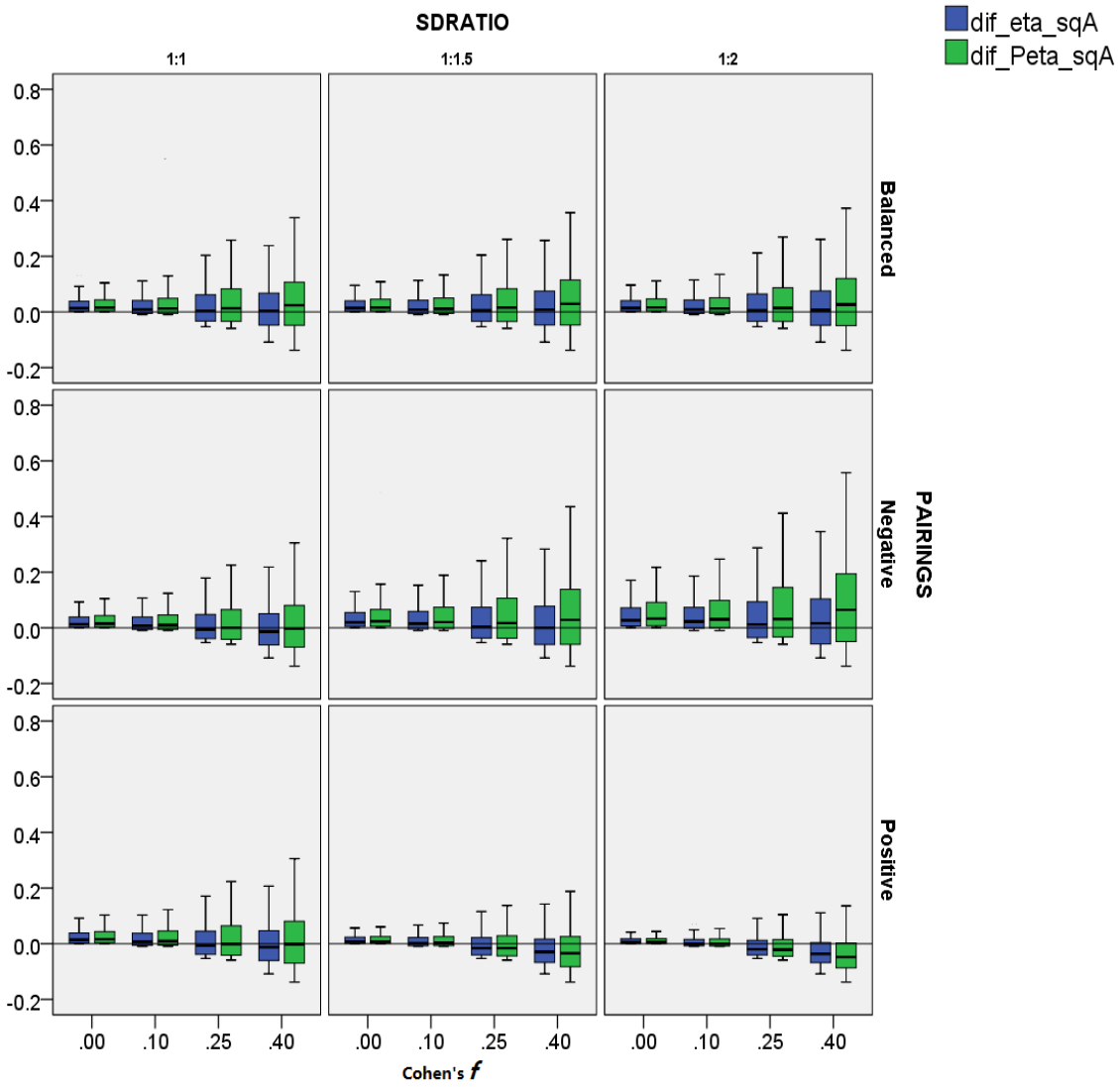

Figure 3. Box-and-Whisker plots for the sampling error bias of $\widehat{\eta}^2$, and $\widehat{\eta}_p^2$ for A-way across heterogeneity, sampling type, and values of Cohen's *f*.

In general, $\hat{\omega}^2$ had the smallest sums of squares, followed by the $\hat{\varepsilon}^2$, and $\hat{\eta}^2$ had the largest sums of squares, which indicated that omega squared yields more accurate parameter estimates. Figure 4 shows the boxplots for bias of $\hat{\eta}^2$, bias of $\hat{\varepsilon}^2$, and bias of $\hat{\omega}^2$ for A-way across heterogeneity, sampling type, and values of Cohen's *f* (the boxplots for the bias of $\hat{\eta}^2$, bias of $\hat{\varepsilon}^2$, and bias of $\hat{\omega}^2$ for the B-way and AB interaction were very close to those depicted in figure 4 and thus are omitted here). As shown in Figure 4, $\hat{\eta}^2$ was slightly inflated, and $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ were more unbiased under a balanced design with homogeneous variance. However, $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ were more affected by the violated homogeneity assumptions and unequal group sizes.

For the A-way, Cohen's *f* accounted for 2%, 0%, 1%, 0%, 1%, and 1% for $\hat{\eta}^2$, $\hat{\varepsilon}^2$, $\hat{\omega}^2$, $\hat{\eta}_p^2$, $\hat{\varepsilon}_p^2$, and $\hat{\omega}_p^2$, respectively. The group size ratios accounted for 3%, 4%, 3%, 4%, 3%, and 4% for the six effect size estimates. The total sample sizes accounted for 4% and 6% for the $\hat{\eta}^2$ and $\hat{\eta}_p^2$, but had minimal effect on the other four types of effect sizes estimates. The interaction between Cohen's *f* and group size ratios accounted for 1%, 1%, 1%, 1%, 1%, and 2% for the six effect size estimates, and the interaction between the standard deviation ratios and group size ratios accounted for 1%, 2%, 2%, 2%, 2%, and 2% for the six effect size estimates. The three-way interaction between Cohen's *f*, standard deviation ratio, and total sample size accounted for 1% for the three partial forms of effect sizes. Effects from all other main and interaction effects were minimal.

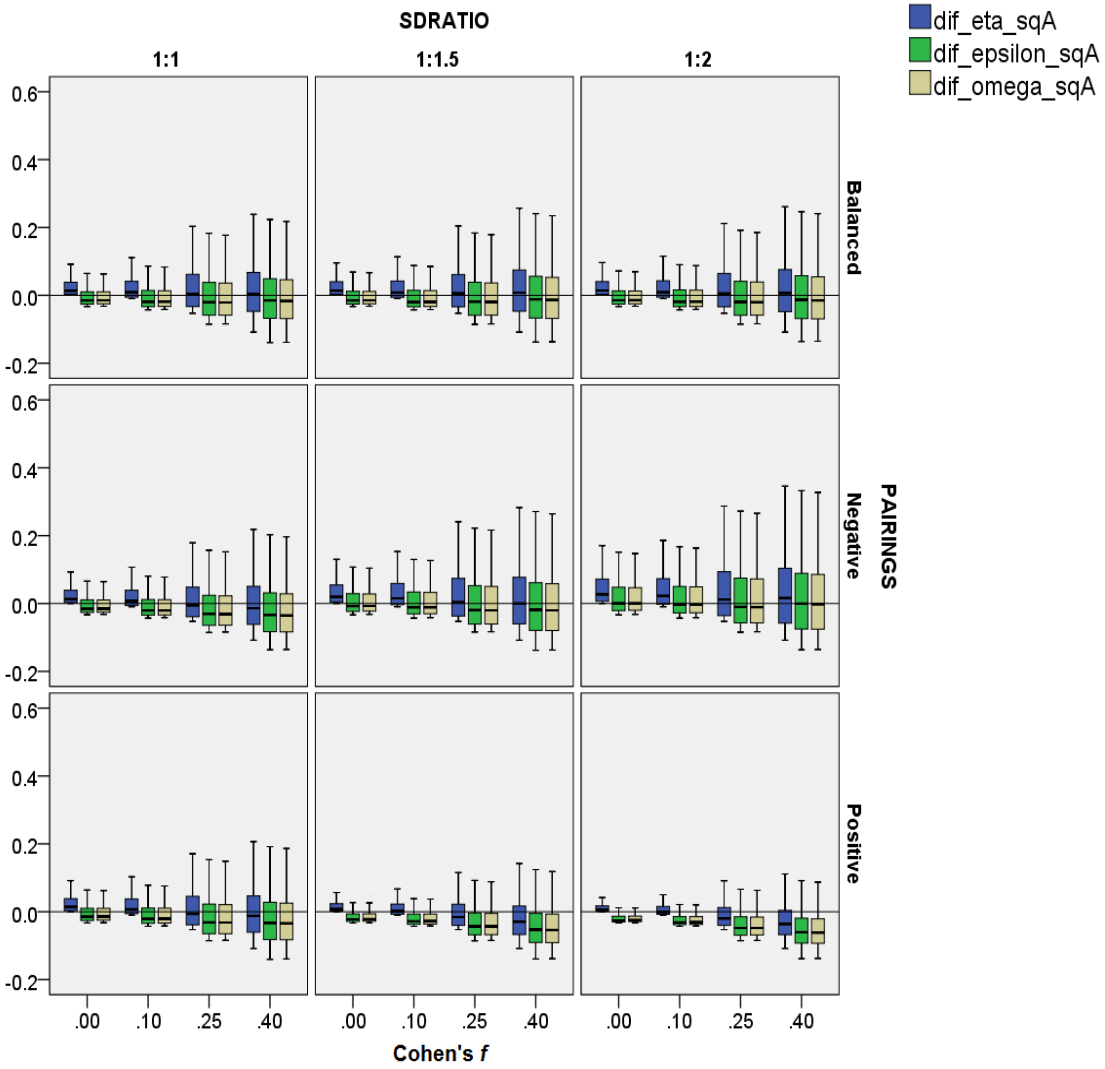

Figure 4. Box-and-Whisker plots for the sampling error bias of $\hat{\eta}^2$, $\hat{\varepsilon}^2$, and $\hat{\omega}^2$ for A-way across heterogeneity, sampling type, and values of Cohen's *f*.

For the B-way, Cohen's *f* accounted for 1% for the $\hat{\eta}^2$. The total sample sizes accounted for 11% and 12% of the bias of $\hat{\eta}^2$ and $\hat{\eta}_p^2$, but the influence of sample sizes on the remaining four forms of effect size estimates were minimal. The group size ratios accounted for 3%, 5%, 5%, 5%, 5%, and 5% for the bias of six forms of effect size estimates. The interaction between Cohen's *f* and group size ratios accounted for 1% of the total bias for the three partial forms: $\hat{\eta}_p^2$, $\hat{\varepsilon}_p^2$, and $\hat{\omega}_p^2$. The interaction between standard deviation ratios and group size ratios accounted for 2%, 3%, 2%, 3%, 2%, and 3% of the total variance, and the three-way interaction between Cohen's *f*, standard deviation, and total sample sizes accounted for 1% for all six forms. The remainder of the main and interaction effect accounted for very little of the total variance in bias.

By comparing the A-way's (two levels) and the B-way's (three levels) results, I found that when the number of levels increased, the Cohen's *f* accounted less for the total variance, but the group size ratios accounted more for the total variance of all six forms of effect size estimates, and the total sample sizes accounted more for the total variance for $\hat{\eta}^2$ and $\hat{\eta}_p^2$.

The AB interaction ANOVA tests had almost identical results as B-way ANOVA tests' results, because the two factors have the same degrees of freedom. Figure 5 shows the boxplots for the sampling error bias of $\hat{\eta}^2$ A-way, B-way, and AB interaction across heterogeneity, sampling type, and values of Cohen's *f* (the boxplots for the sampling error bias of other five forms effect sizes for A-way, B-way, and AB interaction were very similar to those depicted in figure 5, and thus are omitted here). The bias of the

effect size estimates were almost equally "spread out" for the B-way, and the AB interaction. The estimates for the B-way, and the AB interaction were more inflated than the A-way, and were more influenced by the negative pairing, but were less influenced by the positive pairing.

Across the A-way, B-way or AB interaction, the total sample sizes always accounted for the biggest proportion of variance in estimating the $\hat{\eta}^2$ and $\hat{\eta}_p^2$, but not for the other four forms of effect size estimates (i.e., $\hat{\varepsilon}^2$, $\hat{\omega}^2$, $\hat{\varepsilon}_p^2$, and $\hat{\omega}_p^2$). The effect of heterogeneity was enforced with increases in Cohen's *f*, because the interaction between Cohen's *f* and the standard deviation accounted 1-2% of the total variance. And the pairing accounted for 2-3% of the total variance.

Absolute Parameter Bias

The difference between the simple bias and the absolute bias is that simple bias is the difference between the estimated parameters and the true parameters, which can be either positive or negative. A positive bias means the estimated value is greater than the true value while a negative bias means the estimated value is smaller than the true value. The simple bias is a good index for evaluating the **accuracy** of estimates, and the smaller the total variances, the more accurate the estimated results. The absolute bias is the absolute difference between the estimated value and the true value, which is always a positive value. Absolute bias is an index of the **robustness** of the estimates. The smaller the total sums of squares, the more robust the estimates.

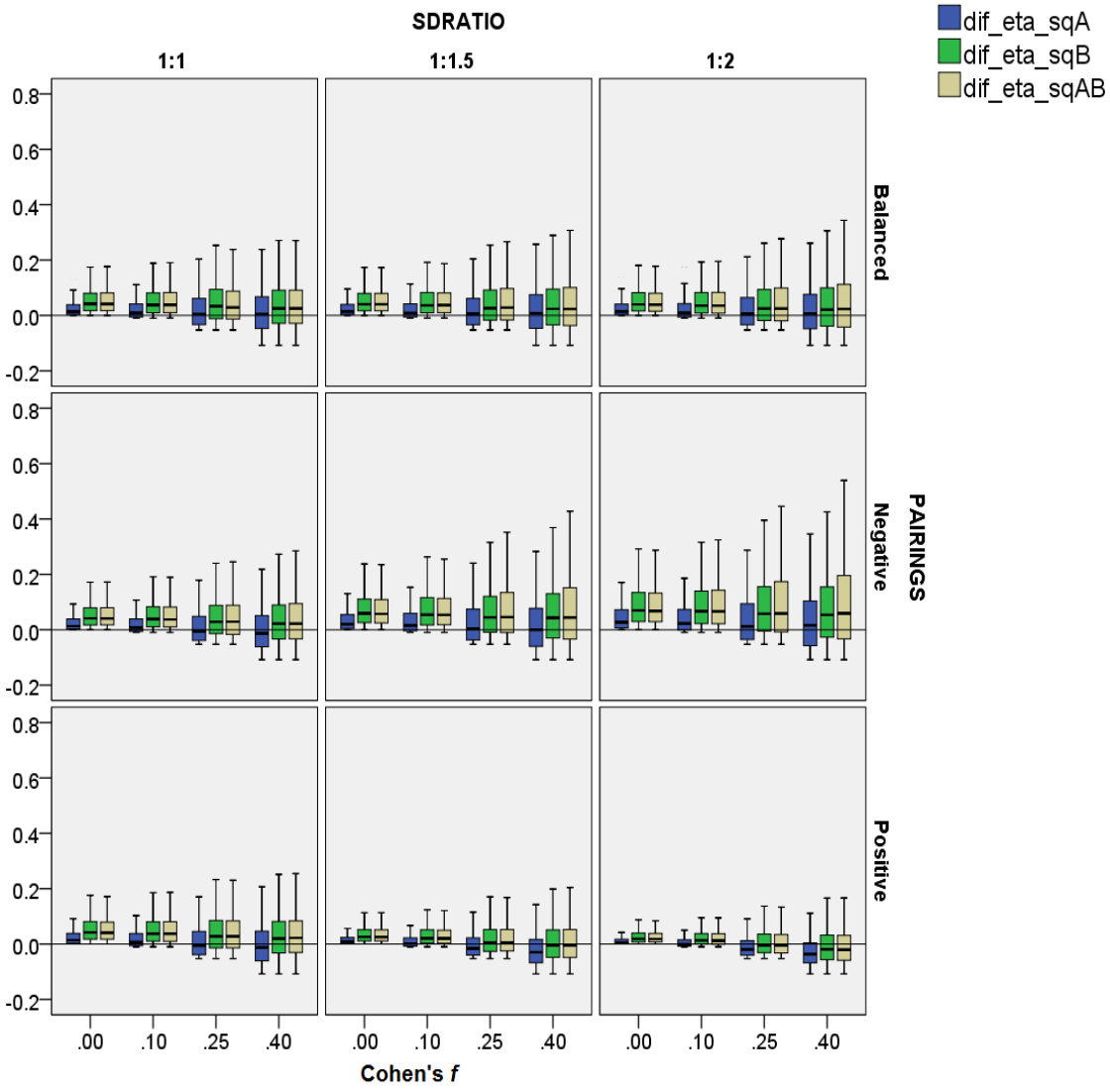

Figure 5. Box-and-Whisker plots for the sampling error bias for A-Way, B-way, and AB interaction across heterogeneity, sampling type, and values of Cohen's *f*.

Table 14, 15, 16 provide the 18 four-way full factorial ANOVA results for absolute bias. The estimation methods were least squares. For the A-way, Cohen's *f* accounted for 10%, 12%, 13%, 15%, 14%, and 18% of the total variance for the six forms of effect sizes (i.e., $\hat{\eta}^2, \hat{\varepsilon}^2, \hat{\omega}^2, \hat{\eta}_p^2, \hat{\varepsilon}_p^2$, and $\hat{\omega}_p^2$), respectively. Total sample sizes accounted for 11%, 11%, 16%, 15%, 16%, and 14% for the six forms of effect sizes, respectively. Group size ratios accounted for 1%, 2%, 1%, 1%, 1%, and 1% of the total variances for the six forms of effect size estimates. The interaction between standard deviation ratios and group size ratios, the interaction between Cohen's *f* and total sample sizes, and the interaction between group size ratios and total sample sizes each accounted for around 1% of the total variance. The remaining main and interaction effects accounted for very little of the total variance in absolute bias.

Just like the results provided in the ANOVA tests on simple bias, the B-way and the AB interaction have very similar results. When the degrees of freedom were increased, the Cohen's *f* accounted for less of the total variance in absolute bias, but the influence of total sample size increased.

Overall, the classical forms of effect sizes (i.e., $\hat{\eta}^2, \hat{\varepsilon}^2$, and $\hat{\omega}^2$) were more robust than the alternative forms of effect sizes (i.e., $\hat{\eta}_p^2$, $\hat{\varepsilon}_p^2$, and $\hat{\omega}_p^2$) when used to estimate the population effect sizes. Omega squared and epsilon squared produced the more robust results, and eta squared was the least robust.

Table 14

Estimated Absolute Bias for A-Way Effect Sizes of η^2, ε^2, ω^2 in the 2 × 3 ANOVA

A-way		η^2		Partial η^2		ε^2		Partial ε^2		ω^2		Partial ω^2	
Source	df	SS	η^2	SS	η^2	SS	η^2	SS	η^2	SS	η^2	SS	η^2
Cohen's *f*	3	123.53	0.10	260.78	0.12	141.76	0.13	289.51	0.15	142.36	0.14	280.64	0.18
SD Ratio	2	1.37	0.00	5.57	0.00	1.98	0.00	6.58	0.00	1.97	0.00	5.88	0.00
Group Size Ratio	2	16.66	0.01	36.73	0.02	6.01	0.01	16.98	0.01	5.69	0.01	11.70	0.01
Total *N*	1	140.35	0.11	251.46	0.11	174.23	0.16	293.48	0.15	165.75	0.16	219.52	0.14
Cohen's *f* * SD Ratio	6	0.50	0.00	2.08	0.00	0.44	0.00	2.14	0.00	0.45	0.00	2.04	0.00
Cohen's f * Group Size Ratio	6	0.38	0.00	0.85	0.00	0.59	0.00	1.56	0.00	0.66	0.00	2.45	0.00
SD Ratio *Group Size Ratio	4	8.71	0.01	23.75	0.01	3.03	0.00	11.17	0.01	2.86	0.00	7.45	0.00
Cohen's f * Total *N*	3	4.93	0.00	16.67	0.01	6.85	0.01	20.89	0.01	6.92	0.01	18.03	0.01
SD Ratio * Total *N*	2	0.03	0.00	0.54	0.00	0.11	0.00	0.80	0.00	0.11	0.00	0.63	0.00
Group Size Ratio * Total *N*	2	9.81	0.01	26.10	0.01	3.66	0.00	13.12	0.01	3.45	0.00	8.91	0.01
Cohen's f * SD Ratio * group Size Ratio	12	0.25	0.00	0.76	0.00	0.22	0.00	0.81	0.00	0.25	0.00	0.77	0.00
Cohen's *f* * SD Ratio * Total *N*	6	0.04	0.00	0.06	0.00	0.05	0.00	0.03	0.00	0.05	0.00	0.02	0.00
Cohen's *f* * Group Size ratio * Total *N*	6	0.76	0.00	3.22	0.00	0.68	0.00	2.75	0.00	0.63	0.00	1.60	0.00
SD Ratio * Group Size Ratio * Total *N*	4	4.87	0.00	15.16	0.01	1.81	0.00	7.76	0.00	1.72	0.00	5.30	0.00
Cohen's f * SD Ratio * Group Size Ratio * Total *N*	12	0.17	0.00	1.24	0.00	0.26	0.00	1.35	0.00	0.24	0.00	0.78	0.00
Error	719928	918.02	0.75	1566.67	0.71	715.10	0.68	1240.51	0.65	689.47	0.67	1004.24	0.64
Total		1230.38		2211.65		1056.80		1909.42		1022.57		1569.97	

Table 15

Estimated Absolute Bias for B-Way Effect Sizes of η^2, ε^2, ω^2 in the 2 × 3 ANOVA

B-way		η^2		Partial η^2		ε^2		Partial ε^2		ω^2		Partial ω^2	
Source	DF	SS	η^2	SS	η^2	SS	η^2	SS	η^2	SS	η^2	SS	η^2
Cohen's *f*	3	51.56	0.02	142.44	0.04	100.51	0.06	223.82	0.09	101.51	0.07	220.10	0.10
SD Ratio	2	3.69	0.00	10.13	0.00	9.59	0.01	19.38	0.01	9.48	0.01	17.84	0.01
Group Size Ratio	2	56.61	0.03	102.10	0.03	14.37	0.01	30.79	0.01	13.74	0.01	23.67	0.01
Total *N*	1	361.36	0.17	568.68	0.17	333.13	0.21	504.87	0.20	316.04	0.21	399.08	0.18
Cohen's *f* * SD Ratio	6	1.43	0.00	3.07	0.00	1.06	0.00	3.11	0.00	1.09	0.00	3.19	0.00
Cohen's f * Group Size Ratio	6	0.12	0.00	0.85	0.00	0.34	0.00	1.13	0.00	0.31	0.00	0.97	0.00
SD Ratio *Group Size Ratio	4	29.50	0.01	61.97	0.02	7.96	0.01	20.82	0.01	7.65	0.01	16.34	0.01
Cohen's f * Total *N*	3	1.09	0.00	2.10	0.00	2.09	0.00	10.93	0.00	2.25	0.00	10.28	0.00
SD Ratio * Total *N*	2	0.06	0.00	0.69	0.00	1.21	0.00	2.97	0.00	1.17	0.00	2.47	0.00
Group Size Ratio * Total *N*	2	19.80	0.01	45.49	0.01	3.54	0.00	12.43	0.00	3.34	0.00	9.06	0.00
Cohen's f * SD Ratio * group Size Ratio	12	0.33	0.00	0.91	0.00	0.17	0.00	0.84	0.00	0.17	0.00	0.67	0.00
Cohen's *f* * SD Ratio * Total *N*	6	0.06	0.00	0.22	0.00	0.11	0.00	0.15	0.00	0.10	0.00	0.12	0.00
Cohen's *f* * Group Size ratio * Total *N*	6	0.34	0.00	0.76	0.00	0.12	0.00	0.63	0.00	0.12	0.00	0.35	0.00
SD Ratio * Group Size Ratio * Total *N*	4	10.31	0.00	26.77	0.01	1.94	0.00	7.90	0.00	1.84	0.00	5.87	0.00
Cohen's f * SD Ratio * Group Size Ratio * Total *N*	12	0.31	0.00	0.22	0.00	0.11	0.00	0.34	0.00	0.11	0.00	0.21	0.00
Error	719928	1608.14	0.75	2414.19	0.71	1100.19	0.70	1704.74	0.67	1061.73	0.70	1454.62	0.67
Total		2144.71		3380.62		1576.43		2544.87		1520.63		2164.86	

Table 16

Estimated Absolute Bias for AB-Interaction Effect Sizes of η^2, ε^2, ω^2 in the 2 × 3 ANOVA

AB Interaction		η^2		Partial η^2		ε^2		Partial ε^2		ω^2		Partial ω^2	
Source	DF	SS	η^2	SS	η^2	SS	η^2	SS	η^2	SS	η^2	SS	η^2
Cohen's *f*	3	74.29	0.03	144.30	0.04	124.56	0.07	223.62	0.09	125.03	0.07	219.56	0.10
SD Ratio	2	8.49	0.00	10.63	0.00	15.66	0.01	20.14	0.01	15.35	0.01	18.47	0.01
Group Size Ratio	2	75.28	0.03	102.38	0.03	22.68	0.01	30.59	0.01	21.67	0.01	23.62	0.01
Total *N*	1	386.64	0.16	565.94	0.17	352.28	0.19	504.27	0.20	333.83	0.19	398.30	0.18
Cohen's *f* * SD Ratio	6	4.94	0.00	3.36	0.00	3.65	0.00	3.31	0.00	3.61	0.00	3.36	0.00
Cohen's f * Group Size Ratio	6	1.88	0.00	0.96	0.00	2.44	0.00	1.30	0.00	2.22	0.00	1.23	0.00
SD Ratio *Group Size Ratio	4	36.83	0.01	63.97	0.02	11.33	0.01	20.99	0.01	10.87	0.01	16.36	0.01
Cohen's f * Total *N*	3	0.86	0.00	2.38	0.00	4.18	0.00	10.98	0.00	4.29	0.00	10.24	0.00
SD Ratio * Total *N*	2	0.72	0.00	0.85	0.00	2.63	0.00	3.46	0.00	2.52	0.00	2.89	0.00
Group Size Ratio * Total *N*	2	25.79	0.01	45.66	0.01	5.46	0.00	11.97	0.00	5.11	0.00	8.70	0.00
Cohen's f * SD Ratio * group Size Ratio	12	0.29	0.00	0.98	0.00	0.60	0.00	0.89	0.00	0.54	0.00	0.63	0.00
Cohen's *f* * SD Ratio * Total *N*	6	0.28	0.00	0.11	0.00	0.09	0.00	0.10	0.00	0.09	0.00	0.08	0.00
Cohen's *f* * Group Size ratio * Total *N*	6	0.12	0.00	0.54	0.00	0.36	0.00	0.50	0.00	0.31	0.00	0.28	0.00
SD Ratio * Group Size Ratio * Total *N*	4	13.03	0.01	28.20	0.01	3.04	0.00	8.28	0.00	2.88	0.00	6.13	0.00
Cohen's f * SD Ratio * Group Size Ratio * Total *N*	12	0.09	0.00	0.30	0.00	0.18	0.00	0.41	0.00	0.15	0.00	0.22	0.00
Error	719928	1834.65	0.74	2405.74	0.71	1273.34	0.70	1699.52	0.67	1227.15	0.70	1450.93	0.67
Total		2464.19		3376.30	1.00	1822.49		2540.32		1755.62		2160.99	

Conclusions

Investigated in the present simulation study were the six forms of effect size estimates. The results from the simulation study answered the following six questions.

Classical Effect Sizes or Partial Alternative Effect Sizes?

Many people believe that the classical forms (i.e., $\hat{\eta}^2$, $\hat{\varepsilon}^2$, and $\hat{\omega}^2$) and the partial alternative forms (i.e., $\hat{\eta}_p^2$, $\hat{\varepsilon}_p^2$, and $\hat{\omega}_p^2$) work equally well in estimating effect sizes because whenever you obtained one result you can simply convert to the other form with the appropriate formula. But the present simulation study demonstrates that the classical forms are more stable and yield fewer fluctuations in estimates even though the means of the bias for the two forms were very close.

Which One Is Better: $\hat{\eta}^2$, $\hat{\varepsilon}^2$, and Partial $\hat{\omega}^2$?

As discovered in previous simulation studies, the $\hat{\eta}^2$ tended to inflate the population effect size estimates, especially when Cohen's *f* was small. And $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ produced less biased more accurate estimates in the balanced design and unbalanced negative pairing conditions, and had a less "spread out" distribution of the bias (e.g., Keselman, 1975). However, $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ are actually more influenced by the "positive pairing" condition than $\hat{\eta}^2$, because the inflation due to sampling error for the $\hat{\eta}^2$ can balance out some of the negative bias in the positive pairing condition.

More importantly, the present simulation study discovered the main factors that cause the bias of estimation on the magnitude of group difference: sampling error, unbalanced design, and types of pairings. Sampling error causes positive bias on $\hat{\eta}^2$, but has a minimal effect on $\hat{\varepsilon}^2$ and $\hat{\omega}^2$. In a balanced design with all assumptions satisfied, $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ are more close to the parameter effect size than $\hat{\eta}^2$ is. An unbalanced group size results in negative bias of the parameter effect size estimate. In a unbalanced design with all assumptions satisfied, $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ do not necessarily yield more accurate estimates than $\hat{\eta}^2$ does, because $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ are deflated due to unequal group sizes but the deflation on $\hat{\eta}^2$ is able to balanced out by the inflation due to sampling error. In the present simulation study with group size ratio equals 1:2, $\hat{\eta}^2$ is more close to the parameter value, and $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ both are negatively biased. Negative pairing causes positive bias for all forms of the effect size estimate, which balances out some negative bias that is present when the group size is not equal. And positive pairing causes negative bias for all forms of the effect size estimate, in which case, $\hat{\eta}^2$ is more likely to have less deflated estimates than $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ will.

The multiplicative impact of the three factors (i.e., sampling error, unbalanced design, and types of pairing) makes the question "which one is better, $\hat{\eta}^2$, $\hat{\varepsilon}^2$, or $\hat{\omega}^2$" not an easy one to answer. But, based on the review study on ANOVA practice, unbalanced designs are more common in educational and psychological research than

balanced designs. Therefore, $\hat{\eta}^2$ is likely to be a better estimate than $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ in certain cases.

What Most Affected the Effect Size Estimates?

The $\hat{\eta}^2$ and $\hat{\eta}_p^2$ were most influenced by the total sample sizes, the smaller the total sample sizes, the more positive bias was observed in the $\hat{\eta}^2$ and $\hat{\eta}_p^2$. But sample sizes have a minimal effect on $\hat{\varepsilon}^2, \hat{\omega}^2$, $\hat{\varepsilon}_p^2$, and $\hat{\omega}_p^2$. When the ANOVA design is unbalanced, the ratio of the group size accounted for a relatively large proportion of the total variance (3-5%). All six forms of the effect size estimates tended to yield negatively biased estimates. However, when small group sizes were paired with large variances, the negative pairing condition balanced out some of the negative bias thus yielding more unbiased estimates. When small group sizes were paired with smaller variances, the positive pairing exacerbated the influence of unbalanced design thus yielding even more negatively biased estimates.

However, as to the absolute bias, the contribution of each main and interaction effect is different. The total sample sizes accounted the greatest proportion of total variance; Cohen's *f* accounted for the second greatest proportion of total variance; and the group size ratios accounted for a small portion of the total variance in bias. All other main and interaction effects had a minimal effect on the absolute bias.

How Does Number of Levels Affect the Effect Size Estimates?

The present study simulated the effect size estimates under a 2 × 3 ANOVA design: the A-way had two levels, the B-way had three levels, and the AB interaction had six levels. The B-way and the AB interaction had very similar results, but the A-way had slightly different results than the B-way and the AB interaction, which indicated that the degrees of freedom were actually the main reason for the differences in effect size estimates. The B-way and the AB interaction had the same number of degree of freedom ($df = 2$) and thus yielded similar results. When the degree of freedom increased for one way, the effect of Cohen's *f* decreased but the effect of group size ratios increased. The degrees of freedom also affected the influence of total sample size on $\hat{\eta}^2$ and $\hat{\eta}_p^2$ estimates; more degrees of freedom resulted in more positive biased on $\hat{\eta}^2$ and $\hat{\eta}_p^2$.

HOW VIOLATION OF ANOVA ASSUMPTIONS IMPACT THE ESTIMATION OF THE INTRACLASS CORRELATION COEFFICIENT FOR A MIXED-EFFECTS MODEL

Criticism of null hypothesis statistical significance testing (NHSST) has dramatically increased in the past few decades. As a result, more and more scholars have strived to find alternatives (e.g., effect sizes and confidence intervals) to supplement NHSST (e.g., Nakagawa & Cuthill, 2007; Schmidt, 1996; Thompson, 1996). At present, editorial policies at more than 20 journals require authors to report estimates of effect sizes (Grissom & Kim, 2012). In the latest American Psychological Association (APA) publication manual, effect sizes have been identified as a necessary element to "convey the most complete meaning of results" (American Psychological Association., 2010, p. 33). More recently, the journal, *Basic and Applied Social Psychology* (*BASP*), has banned *p*-values and confidence intervals, and instead requires authors to provide "strong descriptive statistics, including effect sizes" (Trafimow & Marks, 2015, p. 1). Reform efforts are evidenced also in the increasing number of books, book chapters, and journal articles that discuss appropriate effect sizes for different statistical models (e.g, r^2 for bivariate correlational analysis, R^2 for multiple regression, η^2 for ANOVA). In the popular effect sizes handbook, "Effect Sizes for Research: Univariate and Multivariate Applications", Grissom and Kim (2012) summarized five types of effect sizes for ANOVA: Cohen's *d*, Cohen's *f*, η^2, ε^2, and ω^2. The first two can be interpreted as mean differences, and the last three can be interpreted as the percentage of

explained variance. These are restricted to the fixed-effects model. As a matter of fact, although the random-effects ANOVA is also frequently used in behavioral science, the discussion of effect sizes for random-effects ANOVA is less frequent than for fixed-effects ANOVA (Cardinal & Aitken, 2013).

In the early 1980s, Hedges (1983) proposed a type of effect sizes for random-effects ANOVA, Hedges's *g*, which is analogous to Cohen's *d* and Glass' Δ in the fixed-effects model, and measures the effect in the scale world. Hedges's *g* became an important index in meta- analytical practices (Hedges & Olkin, 2014). But, a more popular effect size for random-effects ANOVA, to date, though still not widely known, is the intraclass correlation coefficient (ICC) (Shrout & Fleiss, 1979).

What is ICC?

Most people know the interclass correlation coefficient – Pearson *r*, which can be used to describe the data structure as paired observations, specifically, (1) to what extent do the two observations have the same order, and (2) to what extent do the two observations have the same shape? (Thompson, 2006). In the interclass correlation coefficient, the two variables are not interchangeable. For example, to calculate the correlation coefficient between intelligence and GPA, scores of intelligence are put in intelligence column, and scores of GPA are put in the GPA column. Switching the places of a pair of intelligence-GPA scores would cause the calculated results to no longer be meaningful.

The other correlation coefficient—the intraclass correlation coefficient (ICC), however, is not as well known the Pearson *r*. ICC is used to describe the pattern within

groups, in other words, how strongly the units clustered within each group. The concept of ICC was originally introduced in social science when cases were interchangeable (e.g., identical twins, competitive siblings, happy and unhappy couples) (Fisher, 1925; Griffin, 1995; Haggard, 1958). For example, in a twins study where two columns contain the values obtained from the twins, it is not required to let the elder twin's scores be filled in the first column and younger twin's scores filled in the second column, or vice versa. Scores in the two columns are interchangeable.

The ICC was developed along with the evolution of sophisticated statistical tools applied to problems in the behavioral sciences (Cook, 2000). The definition of ICC has changed over time. The more recently definition of ICC was framed by the concept of analysis of variance (ANOVA), and more specifically, by the concept of random effects models.

The ICC is generally defined as the variance of interest divided by the total variance, but the exact formula to compute ICC varies in different circumstances. In fact, although there is a variety of ICC statistics, they estimate different population parameters. Shrout and Fleiss (1979) proposed six types of ICCs (see Table 17 for details) and the corresponding circumstances for use. For the same data, different inferential populations would produce remarkably different ICC statistics. Which of the six ICCs is appropriate can be determined by the responding to the following three questions: (1) is the design a one- or two-way ANOVA? (2) Can one effect be ignored in the reliability index? And (3) what is the unit of reliability? Figure 6 provides the specific steps to determine the appropriate ICC.

Table 17

Shrout and Fleiss (1979) Definition of ICCs

Type	One-way random model	Two-way random model	Two-way mixed model
Single	ICC(1,1) $\rho = \sigma_r^2/(\sigma_r^2 + \sigma_e^2)$	ICC(2,1) $\rho = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_c^2 + \sigma_{rc}^2 + \sigma_e^2}$	ICC(3,1) $\rho = \frac{\sigma_r^2 - \sigma_{rc}^2/(k-1)}{\sigma_r^2 + \sigma_c^2 + \sigma_{rc}^2 + \sigma_e^2}$
Average	ICC(1,k) $\rho = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_r^2/k}$	ICC(2,k) $\rho = \frac{\sigma_r^2}{\sigma_r^2 + (\sigma_c^2 + \sigma_{rc}^2 + \sigma_e^2)/k}$	ICC(3,k) $\rho = \frac{\sigma_r^2 - \sigma_{rc}^2/(k-1)}{\sigma_r^2 + (\sigma_c^2 + \sigma_{rc}^2 + \sigma_e^2)/k}$

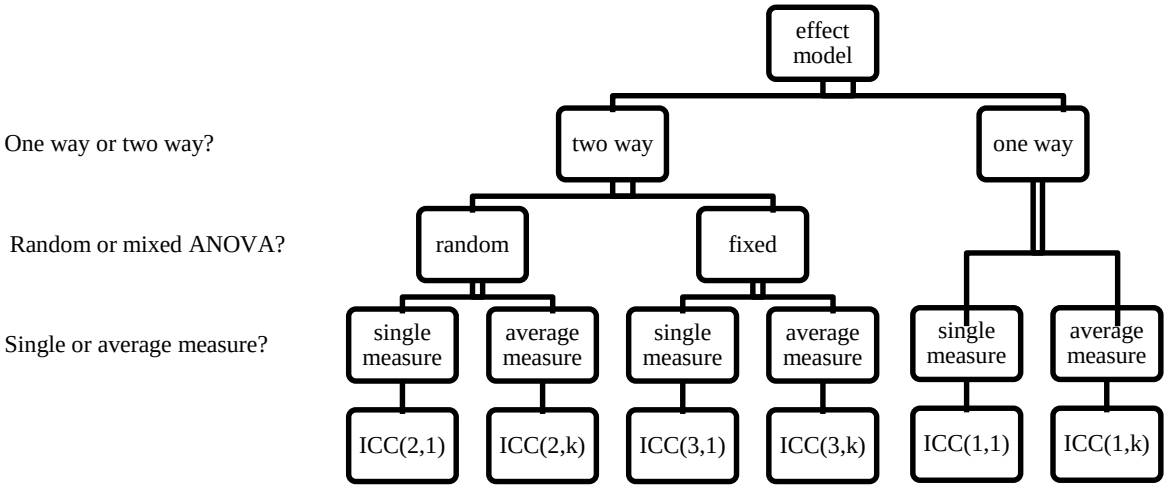

Figure 6. Three questions to determine the appropriate ICC.

Mcgraw and Wong (1996) further extended the number of groups of ICCs. Two more concepts were introduced to differentiate the ICCs: reliability of agreement *vs*. reliability of consistency, and reliability that consider the interaction *vs*. reliability without considering the interaction. Based on Mcgraw and Wong's definitions, the number of types of ICC reached 10 (see Table 18 for details).

Under Mcgraw and Wong's framework, a maximum of five questions were needed to determine the most appropriate type of ICC: (1) is the design a one- or two-way analysis of variance? If this is a one-way model, then ask the second question: (2) what is the unit of reliability? Single case reliability is ICC(1) while group average reliability is ICC(k). If this is a two-way model, then ask the third question: (3) does it have a fixed effect or all random effect? For the random effects model, we need to further determine the type of reliability, which is the fourth question: (4) reliability of consistency or reliability of absolute agreement? And if the model contains a fixed effect, we need to ask the fifth question: (5) does the model contain an interaction effect or not? Figure 7 provides flow chart to determine the most appropriate ICC (this figure is a modified version of Mcgraw and Wong (1996, p. 40) to make this figure more readable in this dissertation).

Table 18

Mcgraw and Wong Definition of ICCs

Type	Single measure	Type	Average measure
ICC(1)	$\rho = \sigma_r^2/(\sigma_r^2 + \sigma_e^2)$	ICC(k)	$\rho = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_r^2/k}$
ICC(2,C,1)	$\rho = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_{rc}^2 + \sigma_e^2}$	ICC(2,C,k)	$\rho = \frac{\sigma_r^2}{\sigma_r^2 + (\sigma_{rc}^2 + \sigma_e^2)/k}$
ICC(2,A,1)	$\rho = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_c^2 + \sigma_{rc}^2 + \sigma_e^2}$	ICC(2,A,k)	$\rho = \frac{\sigma_r^2}{\sigma_r^2 + (\sigma_c^2 + \sigma_{rc}^2 + \sigma_e^2)/k}$
ICC(3,C,1)	$\rho = \frac{\sigma_r^2 - \sigma_{rc}^2/(k-1)}{\sigma_r^2 + \sigma_{rc}^2 + \sigma_e^2}$	ICC(3,C,k)	$\rho = \frac{\sigma_r^2 - \sigma_{rc}^2/(k-1)}{\sigma_r^2 + (\sigma_{rc}^2 + \sigma_e^2)/k}$
ICC(3,A,1)	$\rho = \frac{\sigma_r^2 - \sigma_{rc}^2/(k-1)}{\sigma_r^2 + \sigma_c^2 + \sigma_{rc}^2 + \sigma_e^2}$	ICC(3,A,k)	$\rho = \frac{\sigma_r^2 - \sigma_{rc}^2/(k-1)}{\sigma_r^2 + (\sigma_c^2 + \sigma_{rc}^2 + \sigma_e^2)/k}$

How ICCs Are Applied in Social Sciences

Though the concept of the ICC was considered as early as 1901 when Pearson used a symmetrical correlation table to compute the product-moment ICC for paired scores (Pearson, 1900), and was recommended by Harris as early as 1913 as an effective tool for problems in many fields, such as anthropology, sociology and related fields (Harris, 1913), the use of ICC was rare in the first half of the 20th century (Cook, 2000). Nowadays, the relevance of intraclass correlation to behavioral science has become fairly apparent, but due to the complex computation and complicated definitions, many

behavioral researchers still do not fully understand the concept, not to mention correctly apply the concept in their field of research.

To examine how ICCs were applied currently in social science, I completed a brief systematic review of three peer-reviewed journals: *Journal of Applied Psychology (JAP)*, *Journal of Counseling Psychology (JCP)*, and *Journal of Personality and Social Psychology (JPSP)*. I chose the year 2012 as my target year. A total of 307 articles were downloaded (JAP 92, JCP 61, and JPSP 154). And then, I used the keywords "ICC" or "intraclass correlation" to further screen the target articles, and ended up with a total of 55 articles (JAP 32, JCP 10, and JPSP 13).

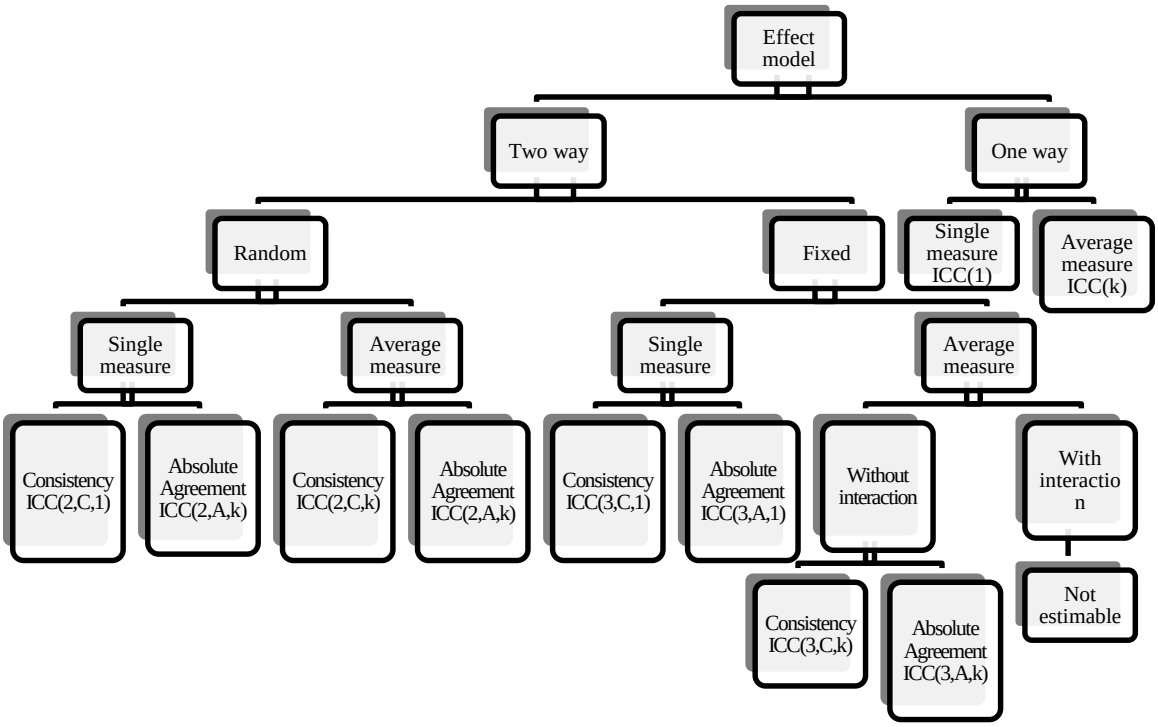

Figure 7. Flowchart to determine the best fit intraclass correlation coefficient (ICC).

Forms of ICC

Three forms of ICCs were reported in the three journals: ICC, ICC(1) and ICC(2), and ICC(j,k). ICC(1) and ICC(2) represent the one-way random single measure of interrater reliability and the one-way random average measure of interrater reliability (Bliese, 2000). ICC(j,k) was created by Shrout and Fleiss (1979), j = (1, 2, or 3): "1" represents a one-way random model, "2" represents a two-way random model, and "3" represents a two-way mixed model; and k = (1, or k): "1" represents single measure, and "k" represents an average measure) (Shrout & Fleiss, 1979).

As shown in Table 19, for all three journals, 49% (n = 27) of the articles reported the intraclass correlation coefficient as "ICC". This form of reporting does not identify which type of ICC was used and thus is likely to cause misuse and misinterpretation of ICC. Bliese's article, "Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis" was well-known in *JAP*. In *JAP*, 72% of the articles reported the ICC as ICC(1) and ICC(2) and cited Bliese (2000) as a reference. Shrout and Fleiss' article "Intraclass correlations: Uses in assessing rater reliability" was more influential in *JPSP*. In *JPSP*, authors tended to report the ICC as ICC(j,k) and cited Shrout and Fleiss (1979) as a reference. But the *JCP* authors did not frequently report the type of ICC used. In 2012, there were ten articles reporting ICC, but none mentioned the specific type used.

Table 19

The Forms of ICCs Reported in JAP, JCP, and JPSP

Journal	ICC(1) & ICC(2)[a]	ICC	ICC(j,k)[b]
JAP	23	8	1
JCP	0	10	0
JPSP	0	9	4

a. The forms of ICCs were proposed by Bliese (2000).
b. The forms of ICCs were proposed by Shrout and Fleiss (1979).

Applications of ICCs

ICC as an index of interrater reliability. Interrater reliability is built under the framework of random effects analysis of variance. The variation between raters is treated either as a random effect or as a fixed effect, based on the assumption of a random condition (e.g., the k raters are randomly selected from a larger population and each rater rates all n targets) or a fixed condition (each target is judged by all the k raters, who are the only research interest) (Shrout & Fleiss, 1979). The ICC is useful in assessing the consistency in ratings among several raters. High ICC means high consensus among raters, and low ICC means low consensus among raters. However, consensus is not equal to accuracy. High ICC does not equate to accurate judgment (Shrout, 1995). In the reviewed articles, interrater reliability was the most popular application of ICC. Twenty articles used ICCs to measure the interrater reliability.

ICC as Aggregation Statistics. Intraclass correlation is one of three statistical techniques used to determine "whether individual level variables constitute sufficient agreement to warrant aggregation or if they are more accurately represented as individuals" (Dixon & Cunningham, 2006, p. 91). Before the multilevel modeling became well known in behavioral sciences, researchers usually simply used the group mean to represent information conveyed from groups, assuming that group means fully represent the whole group, which however, is not always the case. When participants within each group are clustered around the mean for that group, aggregating the data to the group level does not lose much information about the data set; but when participants in the same group are scattered far away from one another, aggregation may distort data dynamics. In practice, to determine if aggregation is acceptable, different researchers choose different cutoff points. Klein et al. (2000) suggested ICC(1) "equal to or above .70 are acceptable, values between .50 and .70 are marginal, and values lower than .50 are poor" (p. 518). Baumgartner, Jackson, Mahar, and Rowe (2003) proposed more stringent criterion n for acceptable reliability: ".70 to .79 is below-average acceptable, .80 to .89 is average acceptable, and .90 to 1.0 is above-average acceptable" (p. 95). In the reviewed articles, 19 articles used ICCs as aggregation statistics.

ICC as descriptive statistics to measure the non-independence. Independence of observations is a required assumption for many statistical methods (e.g., ANOVA, MANOVA, factor analysis). Results are valid only when the independence assumption is reasonably satisfied. The ICC can be used as an index to measure the non-independence of data. A high value of ICC means the cases are not independent – they are correlated

within groups. Five articles in my reviewed articles used the ICCs as the descriptive statistics to measure non-independence.

Some other usage of ICCs. The ICC has many other usages: (1) the ICC can be used to adjust the standard errors when data violate the independence assumption; (2) the ICC can be used as a population parameter to set up simulation conditions, and (3) the ICC can be used to calculate the explained variance between the unconditioned model and the conditioned model in multilevel modeling. Table 20 showed how ICCs were used in the reviewed 54 articles.

Table 20

The Uses of ICC in the Reviewed Articles

Types	Aggregation	Interrater reliability	Adjust standard errors	Non-independence	Others (adjust standard error, simulation condition, partition explained variance, etc.
Counts	19	20	3	5	7

Note. In *JAP*, there was an articles reported the use ICC(1) and ICC(2) but was used to refer another study, thus was excluded in the counting procedure.

What Affects the Estimated ICCs?

Previous Monte Carlo simulation studies have examined the accuracy and robustness of effect sizes for the fixed-effects ANOVA, such as η^2, ε^2, and ω^2 (Keselman, 1975; Skidmore & Thompson, 2013). The results revealed that when all

assumptions were satisfied, ε^2 and ω^2 yielded less biased estimates and η^2 was inflated by sampling error. But when heterogeneity of variance was combined with unequal group sizes, η^2 may yield a better estimate than ε^2 and ω^2 in certain cases. However, the ICC, a well-accepted random effect ANOVA effect size, is rarely examined by quantitative researchers. To the best of my knowledge, no simulation to date has examined the accuracy and robustness of the estimated ICC values based on the random effects ANOVA.

Research Question in the Present Simulation Study

The present simulation study was specifically designed to answer the following two questions: (1) which factors (i.e., population ICC, different effect sizes for the fixed effect, whether or not there was an interaction, ratio of standard deviations, average cell sizes, and type of pairings) affect the accuracy of the estimated parameter ICCs? (2) Which factors (i.e., population ICC, different effect sizes for the fixed effect, whether or not there was an interaction, ratio of standard deviations, average cell sizes, and type of pairings) affect the robustness of the estimated parameter ICCs; and (3) to what extent the estimation of ICC under the framework of analysis of variance is generalizable under different research designs?

Hypothetical Scenario for the Present Simulation Study

Because there are so many types of ICCs and ICCs can be applied in so many circumstances, my simulation study was built on a specific hypothetical scenario. Figure 8 assumed a very common educational scenario, in which the researchers wanted to

examine whether three teaching methods were equally effective for all twelve grades. However, for some reason, the researchers were not able to draw samples from all 12 grades. So they randomly selected two grades from the 12 grades first and then randomly selected samples from the selected grades. The results obtained from the study need to be generalized to all 12 grades.

The Hypothetical scenario was a typical 2 × 3 mixed-effects ANOVA study, in which the course effect was fixed because research interest was limited to the three specific courses, while the grade effect was random because only two grades were selected for study but results were generalized to all grades.

Population Effect Sizes Used in the Simulation

The parameter value for the random effect. An early definition of intraclass correlation only focused on the dyad-level data. In the early 1920s', Fisher (1925) proposed the formula for the ICC for paired data —where the grand mean $\bar{x}$ was defined as $\bar{x} = \frac{1}{2N}\sum_{n=1}^{N}(x_{n,1} + x_{n,2})$, and the total variance s^2 was defined as $s^2 = \frac{1}{2N}\{\sum_{n=1}^{N}(x_{n,1} - \bar{x}) + \sum_{n=1}^{N}(x_{n,2} - \bar{x})\}$, the intraclass correlation $\rho = \frac{1}{Ns^2}\sum_{n=1}^{N}(x_{n,1} - \bar{x})(x_{n,2} - \bar{x})$ (p. 178). In this early definition, ρ could be understood as the portion of the total variance that was contributed between groups. When the group had more than two values, the bivariate covariance, the numerator in the ICC formula, was defined as average covariance for all possible pairs of scores. For example, when the group has three values, $\overline{cov} = \frac{1}{3N}\{\sum_{n=1}^{N}(x_{n,1} - \bar{x})(x_{n,2} - \bar{x}) + \sum_{n=1}^{N}(x_{n,1} - \bar{x})(x_{n,3} - \bar{x}) +$

$\sum_{n=1}^{N}(x_{n,2}-\bar{x})(x_{n,3}-\bar{x})\}$. The earliest definition was an unbiased estimation, and possibly yielded a negative coefficient, but was computably complex.

In the same book, Fisher (1925) proposed another form of the ICC under the framework of analysis of variance. The population ICC $\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\epsilon^2}$, wherein σ_α^2 is variance due to between group difference, and σ_ϵ^2 is the within group variance. This form resolved a difficulty of the early form when group sizes varied between groups, but did not allow the coefficient to be negative.

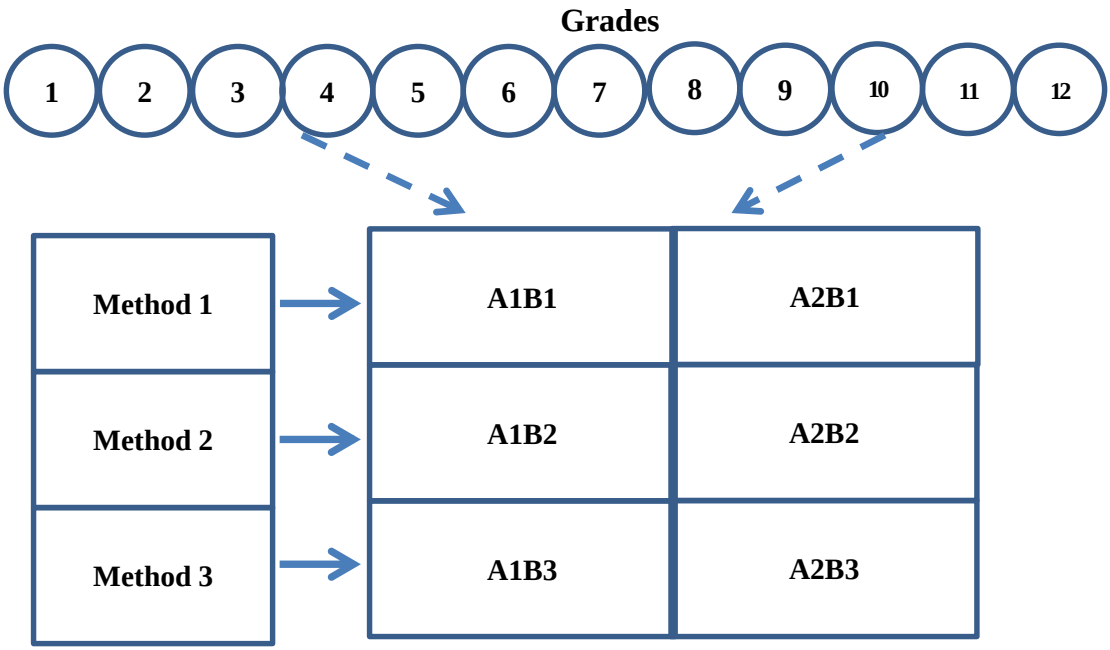

Figure 8. A hypothetical scenario for the simulation study.

In my study, the population for the random effect had 12 values, which yielded $C_{12}^2 = \frac{12\times11}{2\times1}$ = 66 different pairs of scores. For the given set of scores, when a sufficiently large amount of samples (n = 2) were drawn from the population (i.e., the twelve scores), theoretically, each pair of scores would be drawn the same number of times. Therefore, the population between group variance σ_α^2 is the average for all 66 paired scores between group variance.

Table 21 provided 12 normally distributed scores with mean = 0.0, and SD = .99. These scores were generated from the Excel Data Analysis add-in random number generation function. If the twelve scores were used as the means for the twelve grades, the population between group variance σ_α^2 was 0.98812. If k was defined as the multiplier for the between group variance, to control the population ICC, for the given ρ, the $k = \frac{\rho\times\sigma_\epsilon^2}{(1-\rho)\times\sigma_\alpha^2}$. Table 22 provides the k values for different population ICCs and different within group variance σ_ϵ^2. The variation of σ_ϵ^2 was caused by different variance ratios. When the homogeneity of variance was fully satisfied (1:1 for the random effect way, and 1:1:1 for the fixed-effect way), $\sigma_\epsilon^2 = 1$; when the homogeneity of variance was moderately violated, (1:2.25 for the random-effects way, and 1:1: 2.25 for the fix-effects way), $\sigma_\epsilon^2 = \frac{1}{6}(1^2 + 1^2 + 1.5^2 + 1.5^2 + 1.5^2 + 2.25^2) = 2.30208$; and when the homogeneity of variance was severely violated, (1:4 for the random-effects way, and 1:1:4 for the fix-effects way), $\sigma_\epsilon^2 = \frac{1}{6}(1^2 + 1^2 + 1.5^2 + 1.5^2 + 1.5^2 + 2.25^2) = 5$.

Table 21

Normally Distributed Scores (Mean = 0.0, SD =.99, Skewness = .14, Kurtosis = -1.43)

NO.	Value
1	-0.0895079210749827
2	-0.3853574526146990
3	0.3011120952578490
4	-1.3003318599658100
5	-1.0540088624111400
6	0.7882408681325610
7	-0.6051777745597060
8	-1.3114595276420000
9	1.3091130313114300
10	1.3141743693267900
11	-0.2624256012495610
12	1.2661757864407300

Table 22

Parameters Used for Different Population ICCs

ICC ρ	*SD* ratio	σ_ε^2	*k*
0.00	1:1(1:1:1)	1.00000	0.000000000
0.00	1:1.5 (1:1:1.5)	2.30208	0.000000000
0.00	1:2 (1:1:2)	5.00000	0.000000000
0.10	1:1(1:1:1)	1.00000	0.112447240
0.10	1:1.5 (1:1:1.5)	2.30208	0.258862910
0.10	1:2 (1:1:2)	5.00000	0.562236185
0.25	1:1(1:1:1)	1.00000	0.337341710
0.25	1:1.5 (1:1:1.5)	2.30208	0.776588730
0.25	1:2 (1:1:2)	5.00000	1.553177460
0.40	1:1(1:1:1)	1.00000	0.674683420
0.40	1:1.5 (1:1:1.5)	2.30208	1.686708554
0.40	1:2 (1:1:2)	5.00000	3.373417108

Note. Numbers in parentheses reflect the standard deviation ratios for the three levels in the fixed-effect. The three standard deviation ratios for the six cells (SD_{C11}: SD_{C12}:SD_{C13}: SD_{C21}: SD_{C22}: SD_{C23}) were 1: 1: 1: 1: 1: 1, 1: 1: 1.5: 1.5: 1.5: 2.25, 1: 1: 2: 2: 2: 4, respectively.

The parameter value for the fixed-effect and interaction effect. The present simulation study also considered whether or not the magnitude of the fixed-effect and the interaction between the fixed- and random-effects influence the estimated ICC parameters. In the mixed-effects ANOVA, the interaction is automatically treated as random, but with the constraint that $\sum c_j = 0$, $\theta_c^2 = \sum c_j^2/(k-1)$, and $\sum_{j=1}^{k}(rc)_{ij} = 0$ (c is the fixed-effect, and r is the random-effect). To examine whether or not there was a unique contribution from the main effect or the interaction effect, the condition setting considered different size of population fixed-effect with or without the interaction effects. To simplify the condition setting, when there was an interaction, the size of the effect for the interaction was the same as the size of the fixed-effect. Table 23 shows the cell means and cell standard deviations used to control the effect sizes. Three values were used: f = 0, 0.2, and 0.4.

Other Simulation Conditions

Because a previous review study (e.g., Keselman et al., 1998) and my review did not find major differences between the fixed-effects ANOVA and random-effects ANOVA regarding the variance, variance ratio, average cell sizes, and sample size ratio, the present study adopted the same parameters as in the fixed-effect ANOVA simulation: three types of standard deviation ratios (1:1, 1:1.5, and 1:2), two types of average sample sizes (6 and 36), and two types of group size ratios (1:1 and 1:2).

Table 23

Cell Means and Cell Standard Deviations Used for Fixed-Effects' Different Cohen's f, with and without Interaction Effect

Cohen's *f*	SD Ratio	S_{11}	S_{12}	S_{13}	S_{21}	S_{22}	S_{23}	M11	M12	M13	M21	M22	M23
B=0.0, AB=0.0	1:1	1	1	1	1	1	1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
B=0.0, AB=0.0	1:1.5	1	1	1.5	1.5	1.5	2.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
B=0.0, AB=0.0	1:2	1	1	2	2	2	4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
B=0.2, AB=0.0	1:1	1	1	1	1	1	1	-0.2449	0.0000	0.2449	-0.2449	0.0000	0.2449
B=0.2, AB=0.0	1:1.5	1	1	1.5	1.5	1.5	2.25	-0.3717	0.0000	0.3717	-0.3717	0.0000	0.3717
B=0.2, AB=0.0	1:2	1	1	2	2	2	4	-0.5477	0.0000	0.5477	-0.5477	0.0000	0.5477
B=0.4,AB=0.0	1:1	1	1	1	1	1	1	-0.4899	0.0000	0.4899	-0.4899	0.0000	0.4899
B=0.4,AB=0.0	1:1.5	1	1	1.5	1.5	1.5	2.25	-0.7433	0.0000	0.7433	-0.7433	0.0000	0.7433
B=0.4,AB=0.0	1:2	1	1	2	2	2	4	-1.0954	0.0000	1.0954	-1.0954	0.0000	1.0954
B=0.2, AB=0.2	1:1	1	1	1	1	1	1	-0.4899	0.0000	0.4899	0.0000	0.0000	0.0000
B=0.2, AB=0.2	1:1.5	1	1	1.5	1.5	1.5	2.25	-0.7433	0.0000	0.7433	0.0000	0.0000	0.0000
B=0.2, AB=0.2	1:2	1	1	2	2	2	4	-1.0954	0.0000	1.0954	0.0000	0.0000	0.0000
B=0.4,AB=0.4	1:1	1	1	1	1	1	1	-0.9798	0.0000	0.9798	0.0000	0.0000	0.0000
B=0.4,AB=0.4	1:1.5	1	1	1.5	1.5	1.5	2.25	-1.4866	0.0000	1.4866	0.0000	0.0000	0.0000
B=0.4,AB=0.4	1:2	1	1	2	2	2	4	-2.1909	0.0000	2.1909	0.0000	0.0000	0.0000

Note. B represents the fixed-effects, and AB represents the interaction effects.

Replications

This study used 5,000 replications This study in total considered 4 × (3 × 2 - 1) × 3 × 3 × 2 = 360 conditions— four intraclass correlations (k = 0, 0.1, 0.25, and 0.4), three effect sizes for the fixed- and interaction effect (Cohen's f = 0, 0.2, and 0.4), absence of interaction effect (yes or no, when the effect size for the main effect was not zero), three

types of standard deviation ratio (1:1, 1:1.5, and 1:2), two types of average cell sizes (6 and 36), and three types of pairing (balanced, positive pairing, negative pairing).

Simulation Baseline Check

Three indices were chosen to check the simulation baseline: the empirical Type I error rate, the expermentwise error rate, and the power. When the random-effect on the A-way equaled zero, the null model was equivalent to a two-way fixed-effects ANOVA model. Thus the empirical Type I error rate for the A-way should have the same pattern as the B-way and AB-interaction. As reported in the Table 24, when the model assumptions were fully satisfied (equal variance and equal group size), the Type I error rates for A-way, B-way, and AB-interaction were all close to 0.05, and the empirical experimentwise error rates were close to 0.14. As long as the model was balanced, heterogeneity of variance did not have much effect on the Type I error rate, nor did the experimentwise error rate. And as long as the homogeneity assumption was fully satisfied, the unbalanced model did not affect the Type I error rate or experimentwise error rate much. But when an unbalanced model was paired with unequal variance: positive pairing (a large sample size was paired with a large variance) caused inflated Type I error rate and inflated experimentwise error rate; and negative pairing (a large sample size was paired with a small variance) caused deflated Type I error rate and deflated experimentwise error rate. The results were consistent with my fixed-effects ANOVA simulation study and other fixed-effects ANOVA simulation study as well (e.g., Bathke, 2004; Box, 1954; Hsu, 1938; Horsnell, 1953).

Table 24

Empirical Type I Error Rates, Experimentwise Error Rate for A-Way, B-Way, and AB-Interaction, No Random-Effects (Normal Distribution with Average Cell Size Equals Six)

Pairing	SD Ratio	A–way				B–way			AB–Interaction			Empirical Experimentwise error rate
		Random effects (ρ)	Fixed-effects (Cohen's *f*)	Theoretical Type I Error Rate	Empirical Type I Error Rate	Fixed-effects (Cohen's *f*)	Theoretical Type I Error Rate	Empirical Type I Error Rate	Fixed-effects (Cohen's *f*)	Theoretical Type I Error Rate	Empirical Type I Error Rate	
balanced	1:1	0.00	0.00	0.05	0.05	0.00	0.05	0.05	0.00	0.05	0.05	.14
negative pairing	1:1	0.00	0.00	/	0.05	0.00	/	0.05	0.00	/	0.05	.13
positive pairing	1:1	0.00	0.00	/	0.05	0.00	/	0.05	0.00	/	0.05	.13
balanced	1:1.5	0.00	0.00	/	0.05	0.00	0.05	0.05	0.00	0.05	0.05	.14
negative pairing	1:1.5	0.00	0.00	/	0.12	0.00	/	0.13	0.00	/	0.13	.29
positive pairing	1:1.5	0.00	0.00	/	0.01	0.00	/	0.01	0.00	/	0.01	.03
balanced	1:2	0.00	0.00	/	0.06	0.00	0.05	0.07	0.00	0.05	0.06	.15
negative pairing	1:2	0.00	0.00	/	0.18	0.00	/	0.23	0.00	/	0.22	.39
positive pairing	1:2	0.00	0.00	/	0.00	0.00	/	0.01	0.00	/	0.00	.01
balanced	1:1	0.10	0.00	/	0.08	0.00	0.05	0.05	0.00	0.05	0.05	.16
negative pairing	1:1	0.10	0.00	/	0.07	0.00	/	0.05	0.00	/	0.05	.16
positive pairing	1:1	0.10	0.00	/	0.07	0.00	/	0.05	0.00	/	0.05	.16
balanced	1:1.5	0.10	0.00	/	0.11	0.00	0.05	0.05	0.00	0.05	0.06	.19
negative pairing	1:1.5	0.10	0.00	/	0.18	0.00	/	0.13	0.00	/	0.13	.34
positive pairing	1:1.5	0.10	0.00	/	0.04	0.00	/	0.01	0.00	/	0.01	.06
balanced	1:2	0.10	0.00	/	0.19	0.00	0.05	0.07	0.00	0.05	0.07	.27
negative pairing	1:2	0.10	0.00	/	0.30	0.00	/	0.22	0.00	/	0.22	.48
positive pairing	1:2	0.10	0.00	/	0.05	0.00	/	0.01	0.00	/	0.01	.06
balanced	1:1	0.25	0.00	/	0.25	0.00	0.05	0.05	0.00	0.05	0.05	.32
negative pairing	1:1	0.25	0.00	/	0.22	0.00	/	0.05	0.00	/	0.05	.28

Table 24 Continued

Pairing	SD Ratio	A–way				B–way			AB–Interaction			Empirical Experimentwise error rate
		Random effects (ρ)	Fixed-effects (Cohen's *f*)	Theoretical Type I Error Rate	Empirical Type I Error Rate	Fixed-effects (Cohen's *f*)	Theoretical Type I Error Rate	Empirical Type I Error Rate	Fixed-effects (Cohen's *f*)	Theoretical Type I Error Rate	Empirical Type I Error Rate	
positive pairing	1:1	0.25	0.00	/	0.22	0.00	/	0.05	0.00	/	0.04	.28
balanced	1:1.5	0.25	0.00	/	0.43	0.00	0.05	0.05	0.00	0.05	0.06	.48
negative pairing	1:1.5	0.25	0.00	/	0.46	0.00	/	0.13	0.00	/	0.14	.56
positive pairing	1:1.5	0.25	0.00	/	0.30	0.00	/	0.01	0.00	/	0.01	.31
balanced	1:2	0.25	0.00	/	0.59	0.00	0.05	0.08	0.00	0.05	0.07	.64
negative pairing	1:2	0.25	0.00	/	0.64	0.00	/	0.21	0.00	/	0.21	.74
positive pairing	1:2	0.25	0.00	/	0.44	0.00	/	0.00	0.00	/	0.01	.44
balanced	1:1	0.40	0.00	/	0.55	0.00	0.05	0.05	0.00	0.05	0.05	.59
negative pairing	1:1	0.40	0.00	/	0.50	0.00	/	0.05	0.00	/	0.05	.53
positive pairing	1:1	0.40	0.00	/	0.51	0.00	/	0.05	0.00	/	0.05	.56
balanced	1:1.5	0.40	0.00	/	0.69	0.00	0.05	0.05	0.00	0.05	0.06	.72
negative pairing	1:1.5	0.40	0.00	/	0.70	0.00	/	0.14	0.00	/	0.13	.76
positive pairing	1:1.5	0.40	0.00	/	0.62	0.00	/	0.01	0.00	/	0.01	.63
balanced	1:2	0.40	0.00	/	0.79	0.00	0.05	0.07	0.00	0.05	0.07	.81
negative pairing	1:2	0.40	0.00	/	0.81	0.00	/	0.21	0.00	/	0.23	86
positive pairing	1:2	0.40	0.00	/	0.71	0.00	/	0.01	0.00	/	0.01	.72

But when random-effects existed, the Type I error rates for the A-way were inflated by the random-effects: the larger random-effects, the more inflated was the Type I error rate. The existence of random-effects had limited effects on the Type I error rate for the B-way, and AB-interaction. Table 25 provides the empirical Type I error rate for the A-way with the existence of a random-effect. The empirical power for the A-way was very close to the theoretical power that treated the random effects as fixed, which indicated that if a random-effects model was misspecified as a fixed-effects model, researchers were more likely to reject the null while there were actually no fixed effect.

Table 25 also provides the empirical power for the B-way, and AB-interaction. The existence of random-effects had limited effects on the B-way and AB-interaction. The empirical power for the B-way and AB-interaction was very close to the theoretical power when all assumptions were satisfied and was inflated when the sample size and variance were positively paired, and deflated when the sample size and variance were negatively paired. The AB-interaction became random effects when there was a nonzero effect on the A-way, but the change did not affect the empirical power for the AB-interaction.

Table 25

Empirical Type I Error Rate for A-Way, Empirical Power for B-Way, and AB-Interaction, Random-Effects Existed (Normal Distribution with Average Cell Size Equals Six)

Pairing	SD Ratio	A−way				B−way			AB−Interaction		
		Random effects (ρ)	Fixed-effects (Cohen's *f*)	Theoretical Type I error rate	Empirical Type I error rate	Fixed-effects (Cohen's *f*)	Theoretical Power	Empirical Power	Fixed-effects (Cohen's *f*)	Theoretical Power	Empirical Power
balanced	1:1	0.10	0.00	/	0.08	0.20	0.16	0.16	0.20	0.16	0.16
negative pairing	1:1	0.10	0.00	/	0.07	0.20	/	0.15	0.20	/	0.15
positive pairing	1:1	0.10	0.00	/	0.07	0.20	/	0.15	0.20	/	0.16
balanced	1:1	0.10	0.00	/	0.07	0.40	0.52	0.51	0.40	0.52	0.52
negative pairing	1:1	0.10	0.00	/	0.07	0.40	/	0.48	0.40	/	0.48
positive pairing	1:1	0.10	0.00	/	0.06	0.40	/	0.48	0.40	/	0.48
balanced	1:1.5	0.10	0.00	/	0.12	0.20	0.16	0.17	0.20	0.16	0.17
negative pairing	1:1.5	0.10	0.00	/	0.17	0.20	/	0.27	0.20	/	0.27
positive pairing	1:1.5	0.10	0.00	/	0.04	0.20	/	0.06	0.20	/	0.06
balanced	1:1.5	0.10	0.00	/	0.11	0.40	0.52	0.51	0.40	0.52	0.50
negative pairing	1:1.5	0.10	0.00	/	0.17	0.40	/	0.59	0.40	/	0.61
positive pairing	1:1.5	0.10	0.00	/	0.04	0.40	/	0.31	0.40	/	0.31
balanced	1:2	0.10	0.00	/	0.19	0.20	0.16	0.17	0.20	0.16	0.17
negative pairing	1:2	0.10	0.00	/	0.30	0.20	/	0.35	0.20	/	0.35
positive pairing	1:2	0.10	0.00	/	0.04	0.20	/	0.04	0.20	/	0.03
balanced	1:2	0.10	0.00	/	0.18	0.40	0.52	0.50	0.40	0.52	0.51
negative pairing	1:2	0.10	0.00	/	0.29	0.40	/	0.66	0.40	/	0.65
positive pairing	1:2	0.10	0.00	/	0.05	0.40	/	0.22	0.40	/	0.23
balanced	1:1	0.25	0.00	/	0.27	0.20	0.16	0.15	0.20	0.16	0.16
negative pairing	1:1	0.25	0.00	/	0.21	0.20	/	0.14	0.20	/	0.15

Table 25 Continued

Pairing	SD Ratio	A−way				B−way			AB−Interaction		
		Random effects (ρ)	Fixed-effects (Cohen's *f*)	Theoretical Type I error rate	Empirical Type I error rate	Fixed-effects (Cohen's *f*)	Theoretical Power	Empirical Power	Fixed-effects (Cohen's *f*)	Theoretical Power	Empirical Power
positive pairing	1:1	0.25	0.00	/	0.23	0.20	/	0.15	0.20	/	0.15
balanced	1:1	0.25	0.00	/	0.27	0.40	0.52	0.52	0.40	0.52	0.53
negative pairing	1:1	0.25	0.00	/	0.23	0.40	/	0.48	0.40	/	0.48
positive pairing	1:1	0.25	0.00	/	0.22	0.40	/	0.50	0.40	/	0.49
balanced	1:1.5	0.25	0.00	/	0.44	0.20	0.16	0.16	0.20	0.16	0.16
negative pairing	1:1.5	0.25	0.00	/	0.45	0.20	/	0.26	0.20	/	0.27
positive pairing	1:1.5	0.25	0.00	/	0.29	0.20	/	0.06	0.20	/	0.06
balanced	1:1.5	0.25	0.00	/	0.44	0.40	0.52	0.51	0.40	0.52	0.52
negative pairing	1:1.5	0.25	0.00	/	0.47	0.40	/	0.60	0.40	/	0.59
positive pairing	1:1.5	0.25	0.00	/	0.30	0.40	/	0.32	0.40	/	0.31
balanced	1:2	0.25	0.00	/	0.60	0.20	0.16	0.18	0.20	0.16	0.18
negative pairing	1:2	0.25	0.00	/	0.65	0.20	/	0.34	0.20	/	0.36
positive pairing	1:2	0.25	0.00	/	0.44	0.20	/	0.03	0.20	/	0.04
balanced	1:2	0.25	0.00	/	0.60	0.40	0.52	0.51	0.40	0.52	0.51
negative pairing	1:2	0.25	0.00	/	0.65	0.40	/	0.65	0.40	/	0.66
positive pairing	1:2	0.25	0.00	/	0.44	0.40	/	0.22	0.40	/	0.23
balanced	1:1	0.40	0.00	/	0.54	0.20	0.16	0.16	0.20	0.16	0.16
negative pairing	1:1	0.40	0.00	/	0.50	0.20	/	0.15	0.20	/	0.15
positive pairing	1:1	0.40	0.00	/	0.51	0.20	/	0.15	0.20	/	0.15
balanced	1:1	0.40	0.00	/	0.55	0.40	0.52	0.52	0.40	0.52	0.52
negative pairing	1:1	0.40	0.00	/	0.50	0.40	/	0.47	0.40	/	0.49
positive pairing	1:1	0.40	0.00	/	0.49	0.40	/	0.49	0.40	/	0.49
balanced	1:1.5	0.40	0.00	/	0.71	0.20	0.16	0.17	0.20	0.16	0.18

Table 25 Continued

Pairing	SD Ratio	A−way				B−way			AB−Interaction		
		Random effects (ρ)	Fixed-effects (Cohen's *f*)	Theoretical Type I error rate	Empirical Type I error rate	Fixed-effects (Cohen's *f*)	Theoretical Power	Empirical Power	Fixed-effects (Cohen's *f*)	Theoretical Power	Empirical Power
negative pairing	1:1.5	0.40	0.00	/	0.70	0.20	/	0.28	0.20	/	0.27
positive pairing	1:1.5	0.40	0.00	/	0.61	0.20	/	0.06	0.20	/	0.06
balanced	1:1.5	0.40	0.00	/	0.70	0.40	0.52	0.51	0.40	0.52	0.51
negative pairing	1:1.5	0.40	0.00	/	0.71	0.40	/	0.60	0.40	/	0.60
positive pairing	1:1.5	0.40	0.00	/	0.62	0.40	/	0.32	0.40	/	0.31
balanced	1:2	0.4	0.00	/	0.79	0.20	0.16	0.17	0.20	0.16	0.17
negative pairing	1:2	0.4	0.00	/	0.81	0.20	/	0.36	0.20	/	0.34
positive pairing	1:2	0.4	0.00	/	0.71	0.20	/	0.03	0.20	/	0.04
balanced	1:2	0.4	0.00	/	0.79	0.40	0.52	0.51	0.40	0.52	0.52
negative pairing	1:2	0.4	0.00	/	0.82	0.40	/	0.65	0.40	/	0.65
positive pairing	1:2	0.4	0.00	/	0.71	0.40	/	0.22	0.40	/	0.22

Results

The mixed-effect ANOVA simulation study obtained 1,800,000 independent samples under 360 conditions (four intraclass correlation (k = 0, 0.1, 0.25, and 0.4), three effect sizes for the fixed- and interaction effects (Cohen's f = 0, 0.2, and 0.4), absence of interaction effects (yes or no, when the effect size for the main effect was not zero), three standard deviation ratios (1:1, 1:1.5, and 1:2), two average cell sizes (6 and 36), and three types of pairings (balanced, positive pairing, negative pairing). For each sample, the estimated ICC was calculated with the formula $ICC(3,1) = \frac{MSB-MSW}{MSB+(k-1)*MSW}$ (MSB is the estimated mean square for the A-way, and MSW is the estimated mean square for the error, and k is the number of levels for A-way). The bias (i.e., the difference between the estimated ICC and the population ICC) and absolute bias (i.e., the absolute difference between the estimated ICC and the population ICC) of the estimated ICC was then calculated, and a full factorial ANOVA with the five conditions (i.e., population ICC for the A-way, population Cohen's f for the B-way, whether or not the two ways have an interaction effect, variance ratio, group size ratio, and average cell sizes) was conducted to further explore which condition had a greater effect on the estimation of ICCs.

Parameter Bias

Provided in Table 26 is the five-way full factorial ANOVA with the least squares estimation methods for the estimated parameter bias for the A-way effect. The population ICC yielded the largest effect, which explained 12.6% of the total variance.

Table 26

Estimated Parameter Bias for A-Way ICCs

Source	Type III SS	*df*	η^2
SD Ratio	11561.1	2	0.017
Group Size Ratio	6632.7	2	0.010
Total *N*	21652.7	1	0.031
ρ	87376.7	3	0.126
Cohen's *f*	0.9	2	0.000
Interaction	0.0	1	0.000
SD Ratio * Group Size Ratio	3458.3	4	0.005
SD Ratio * Total *N*	36.3	2	0.000
SD Ratio * ρ	4697.8	6	0.007
SD Ratio * Cohen's *f*	1.0	4	0.000
SD Ratio * Interaction	2.1	2	0.000
Group Size Ratio * Total *N*	252.4	2	0.000
Group Size Ratio * ρ	1686.4	6	0.002
Group Size Ratio * Cohen's *f*	0.9	4	0.000
Group Size Ratio * Interaction	0.0	2	0.000
Total *N* * ρ	8982.1	3	0.013
Total *N* * Cohen's *f*	0.9	2	0.000
Total *N* * Interaction	0.1	1	0.000
ρ * Cohen's *f*	3.6	6	0.000
ρ * Interaction	1.7	3	0.000
Cohen's *f* * Interaction	0.0	1	0.000
SD Ratio * Group Size Ratio * Total *N*	128.4	4	0.000
SD Ratio * Group Size Ratio * ρ	1068.2	12	0.002
SD Ratio * Group Size Ratio * Cohen's *f*	0.9	8	0.000
SD Ratio * Group Size Ratio * Interaction	0.7	4	0.000
SD Ratio * Total *N* * ρ	1329.7	6	0.002
SD Ratio * Total *N* * Cohen's *f*	0.8	4	0.000
SD Ratio * Total *N* * Interaction	0.3	2	0.000
SD Ratio * ρ * Cohen's *f*	3.3	12	0.000
SD Ratio * ρ * Interaction	0.9	6	0.000
SD Ratio * Cohen's *f* * Interaction	0.2	2	0.000
Group Size Ratio * Total *N* * ρ	90.5	6	0.000
Group Size Ratio * Total *N* * Cohen's *f*	0.8	4	0.000
Group Size Ratio * Total *N* * Interaction	0.4	2	0.000
Group Size Ratio * ρ * Cohen's *f*	5.1	12	0.000
Group Size Ratio * ρ * Interaction	3.6	6	0.000

Table 26 Continued

Source	Type III SS	*df*	η^2
Group Size Ratio * Cohen's *f* * Interaction	0.0	2	0.000
Total *N* * ρ * Cohen's *f*	1.6	6	0.000
Total *N* * ρ * Interaction	1.0	3	0.000
Total *N* * Cohen's *f* * Interaction	0.1	1	0.000
ρ * Cohen's *f* * Interaction	2.5	3	0.000
SD Ratio * Group Size Ratio * Total *N* * ρ	37.5	12	0.000
SD Ratio * Group Size Ratio * Total *N* * Cohen's *f*	1.8	8	0.000
SD Ratio * Group Size Ratio * Total *N* * Interaction	0.4	4	0.000
SD Ratio * Group Size Ratio * ρ * Cohen's *f*	5.2	24	0.000
SD Ratio * Group Size Ratio * ρ * Interaction	3.8	12	0.000
SD Ratio * Group Size Ratio * Cohen's *f* * Interaction	1.0	4	0.000
SD Ratio * Total *N* * ρ * Cohen's *f*	4.7	12	0.000
SD Ratio * Total *N* * ρ * Interaction	2.2	6	0.000
SD Ratio * Total *N* * Cohen's *f* * Interaction	0.1	2	0.000
SD Ratio * ρ * Cohen's *f* * Interaction	1.0	6	0.000
Group Size Ratio * Total *N* * ρ * Cohen's *f*	4.2	12	0.000
Group Size Ratio * Total *N* * ρ * Interaction	0.8	6	0.000
Group Size Ratio * Total *N* * Cohen's *f* * Interaction	0.6	2	0.000
Group Size Ratio * ρ * Cohen's *f* * Interaction	1.1	6	0.000
Total N * ρ * Cohen's *f* * Interaction	0.1	3	0.000
SD Ratio * Group Size Ratio * Total *N* * ρ * Cohen's *f*	3.9	24	0.000
SD Ratio * Group Size Ratio * Total *N* * ρ * Interaction	1.9	12	0.000
SD Ratio * Group Size Ratio * Total *N* * Cohen's *f* * Interaction	2.3	4	0.000
SD Ratio * Group Size Ratio * ρ * Cohen's *f* * Interaction	2.1	12	0.000
SD Ratio * Total *N* * ρ * Cohen's *f* * Interaction	1.2	6	0.000
Group Size Ratio * Total *N* * ρ * Cohen's *f* * Interaction	0.8	6	0.000
SD Ratio * Group Size Ratio * Total *N* * ρ * Cohen's *f* * Interaction	5.7	12	0.000
Error	536265.6	1799640	0.773
Total	693955.4	1799999	

The average cell size was the second largest effect, which explained 3.1% of the total variance. The ratio of the group standard deviations explained 1.7% of the total variance. And the ratio of group sizes explained 1% of the total variance. Some two-interactions between the five conditions also explained small percentages of total variance. For example, the interaction between the population ICC and the average cell sizes explained 1.3% of the total variance, the interaction between the ratio of standard deviations and the population ICCs explained 0.7% of the total variance, and the interaction between the ratio of standard deviations and the ratio of group sizes explained 0.5% of the total variance. The effects of all other interactions were minimal.

Absolute Parameter Bias

Provided in Table 27 are the five-way full factorial ANOVA results using least squares estimation methods that analyze the effects that contributed to the absolute parameter bias for the A-way effect. It is interesting to find that the absolute parameter bias (in other words, the robustness of the estimates) was pretty stable under the various study conditions. The ratio of standard deviation explained 0.1% of the total variance, the size of the population ICCs explained 0.1% of the total variance. The effect of all other main effects and interaction effects were minimal.

Table 27

Estimated Absolute Parameter Bias for A-Way ICCs

Source	Type III SS	*df*	$\eta 2$
SD Ratio	367.948	2	0.002
Group Size Ratio	2.576	2	0.000
Total *N*	264.921	1	0.002
ρ	421.946	3	0.003
Cohen's *f*	.116	2	0.000
Interaction	.002	1	0.000
SD Ratio * Group Size Ratio	7.287	4	0.000
SD Ratio * Total *N*	89.383	2	0.001
SD Ratio * ρ	32.405	6	0.000
SD Ratio * Cohen's *f*	.435	4	0.000
SD Ratio * Interaction	.622	2	0.000
Group Size Ratio * Total *N*	16.761	2	0.000
Group Size Ratio * ρ	221.258	6	0.001
Group Size Ratio * Cohen's *f*	.347	4	0.000
Group Size Ratio * Interaction	.623	2	0.000
Total *N* * ρ	335.067	3	0.002
Total *N* * Cohen's *f*	.094	2	0.000
Total *N* * Interaction	.092	1	0.000
ρ * Cohen's *f*	.823	6	0.000
ρ * Interaction	.196	3	0.000
Cohen's *f* * Interaction	.035	1	0.000
SD Ratio * Group Size Ratio * Total *N*	6.074	4	0.000
SD Ratio * Group Size Ratio * ρ	163.108	12	0.001
SD Ratio * Group Size Ratio * Cohen's *f*	.782	8	0.000
SD Ratio * Group Size Ratio * Interaction	.250	4	0.000
SD Ratio * Total *N* * ρ	84.899	6	0.001
SD Ratio * Total *N* * Cohen's *f*	.041	4	0.000
SD Ratio * Total *N* * Interaction	.154	2	0.000
SD Ratio * ρ * Cohen's *f*	.929	12	0.000
SD Ratio * ρ * Interaction	.368	6	0.000
SD Ratio * Cohen's *f* * Interaction	.129	2	0.000
Group Size Ratio * Total *N* * ρ	45.286	6	0.000
Group Size Ratio * Total *N* * Cohen's *f*	.127	4	0.000
Group Size Ratio * Total *N* * Interaction	.015	2	0.000

Table 27 Continued

Source	Type III SS	*df*	$\eta 2$
Group Size Ratio * ρ * Cohen's *f*	.971	12	0.000
Group Size Ratio * ρ * Interaction	.800	6	0.000
Group Size Ratio * Cohen's *f* * Interaction	.100	2	0.000
Total *N* * ρ * Cohen's *f*	1.005	6	0.000
Total *N* * ρ * Interaction	.218	3	0.000
Total *N* * Cohen's *f* * Interaction	.046	1	0.000
ρ * Cohen's *f* * Interaction	.067	3	0.000
SD Ratio * Group Size Ratio * Total *N* * ρ	31.159	12	0.000
SD Ratio * Group Size Ratio * Total *N* * Cohen's *f*	.830	8	0.000
SD Ratio * Group Size Ratio * Total *N* * Interaction	.454	4	0.000
SD Ratio * Group Size Ratio * ρ * Cohen's *f*	1.581	24	0.000
SD Ratio * Group Size Ratio * ρ * Interaction	.570	12	0.000
SD Ratio * Group Size Ratio * Cohen's *f* * Interaction	.252	4	0.000
SD Ratio * Total *N* * ρ * Cohen's *f*	1.697	12	0.000
SD Ratio * Total *N* * ρ * Interaction	1.481	6	0.000
SD Ratio * Total *N* * Cohen's *f* * Interaction	.392	2	0.000
SD Ratio * ρ * Cohen's *f* * Interaction	.463	6	0.000
Group Size Ratio * Total *N* * ρ * Cohen's *f*	.701	12	0.000
Group Size Ratio * Total *N* * ρ * Interaction	.239	6	0.000
Group Size Ratio * Total *N* * Cohen's *f* * Interaction	.048	2	0.000
Group Size Ratio * ρ * Cohen's *f* * Interaction	.269	6	0.000
Total N * ρ * Cohen's *f* * Interaction	.076	3	0.000
SD Ratio * Group Size Ratio * Total *N* * ρ * Cohen's *f*	1.330	24	0.000
SD Ratio * Group Size Ratio * Total *N* * ρ * Interaction	.572	12	0.000
SD Ratio * Group Size Ratio * Total *N* * Cohen's *f* * Interaction	.654	4	0.000
SD Ratio * Group Size Ratio * ρ * Cohen's *f* * Interaction	1.341	12	0.000
SD Ratio * Total *N* * ρ * Cohen's *f* * Interaction	.206	6	0.000
Group Size Ratio * Total *N* * ρ * Cohen's *f* * Interaction	.261	6	0.000
SD Ratio * Group Size Ratio * Total *N* * ρ * Cohen's *f* * Interaction	1.197	12	0.000
Error	151151.476	1799640	0.985
Total	153392.7	1799999	1.000

Discussion

The estimate of ICCs in the mixed-effect ANOVA model is a more complicated case compared with the estimates of effect sizes in the fixed-effects ANOVA. The present simulation study examined five possible effects (i.e., population ICC, different effect sizes for the fixed-effect, whether or not there was an interaction, ratio of standard deviations, average cell sizes, and type of pairings) that impact the accuracy and robustness of the estimate of population ICCs. By analyzing the 1,8000,000 independent samples (four intraclass correlation (k = 0, 0.1, 0.25, and 0.4), three effect sizes for the fixed- and interaction effect (Cohen's f = 0, 0.2, and 0.4), absence of interaction effect (yes or no, when the effect size for the main effect was not zero), three types of standard deviation ratios (1:1, 1:1.5, and 1:2), two types of average cell sizes (6 and 36), and three types of pairings (balanced, positive pairing, negative pairing), three questions could be addressed.

What Affects the Accuracy of the Estimated Parameter ICCs?

Provided in Figure 9 are box-and-whisker plots for the sampling error bias of estimated ICC values across heterogeneous variances, sampling types and sizes of population ICCs. As shown in the figure, the bias was not consistently negative or positive and was affected mostly by the population ICCs. For example, when all assumptions were fully satisfied (i.e., fully homogeneous variances, equal group size, normal distributions, and independence of data), the population ICCs were likely to be underestimated when the population ICCs were small (e.g., ρ = 0 or ρ = 0.1),

overestimated when the population ICCs were large (e.g., $\rho = 0.4$), and more accurately estimated when the population ICCs were of moderate size (e.g., $\rho = 0.25$).

When the model was balanced, heterogeneity of variance inflated the estimation of ICCs; when group sizes were not equal, unequal group sizes did not affect the population ICC values. But when small group sizes were paired with large variances, the negative pairing inflated the parameter estimates. However, the positive pairing (i.e., small group size was paired with small variances) did not consistently inflate or deflate the parameter estimates. Whether or not the estimates were inflated or deflated was largely dependent on the size of the population ICCs and the extent to which the homogeneity assumption was violated. For example, during the positive pairing, a moderate violation of the homogeneity assumption deflated the population estimate when the population ICC equaled 0.1, but when the violation of homogeneity assumption became severe, the inflation no longer occurred.

The accuracy of estimation was also affected by average cell sizes. Depicted in Figure 10 are the box-and-whisker plots for the sampling error bias of estimated ICCs for the A-way across heterogeneity type, sampling type, and cell sizes ($n = 6$, and $n = 36$). As showing in Figure 10, small average cell sizes yielded more accurate estimates of parameter ICCs in all circumstances (e.g., heterogeneity type, group size condition, and pairing type).

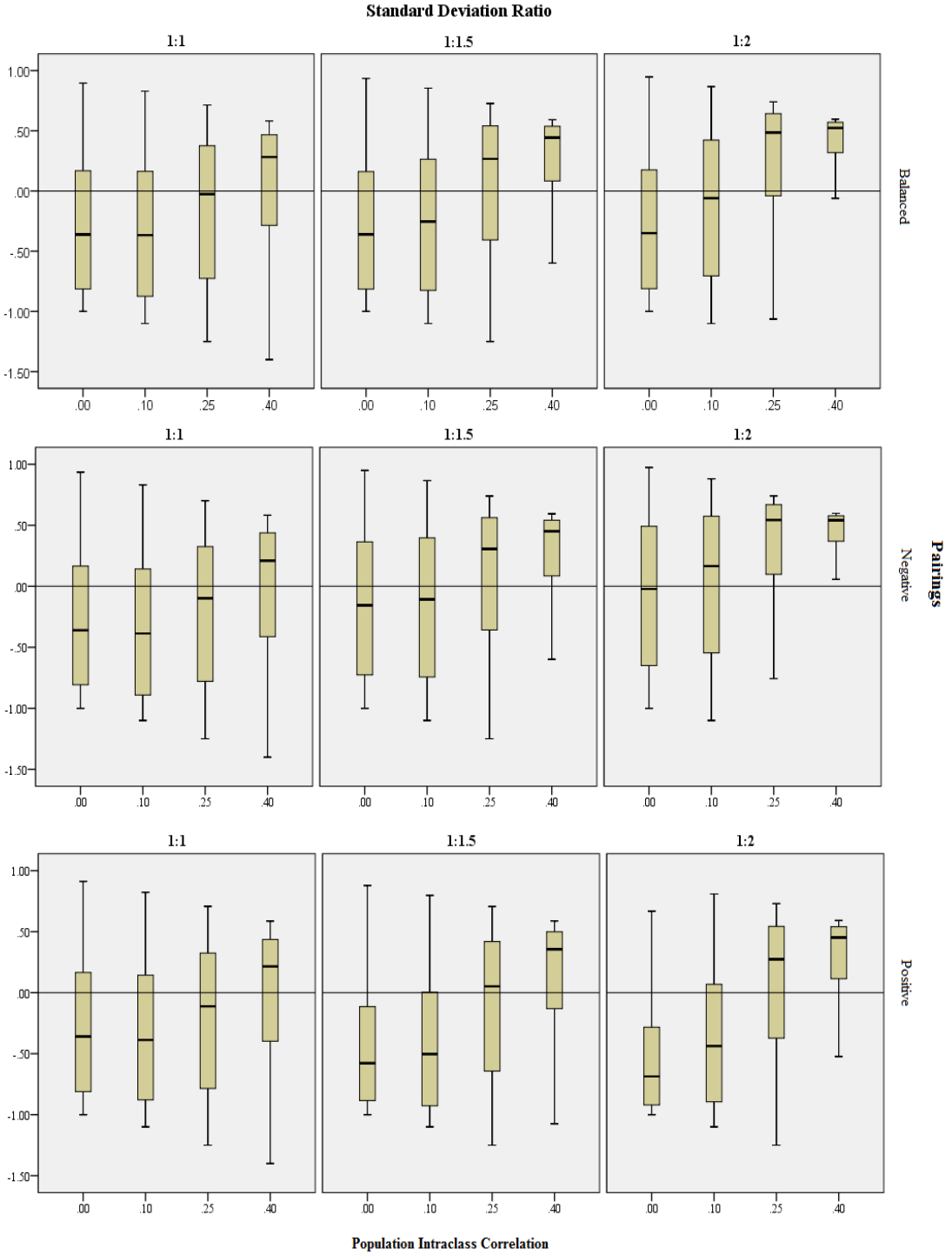

Figure 9. Box-and-Whisker plots for the sampling error bias of estimated ICC for the A-way across heterogeneity, sampling type, and size of population ICC.

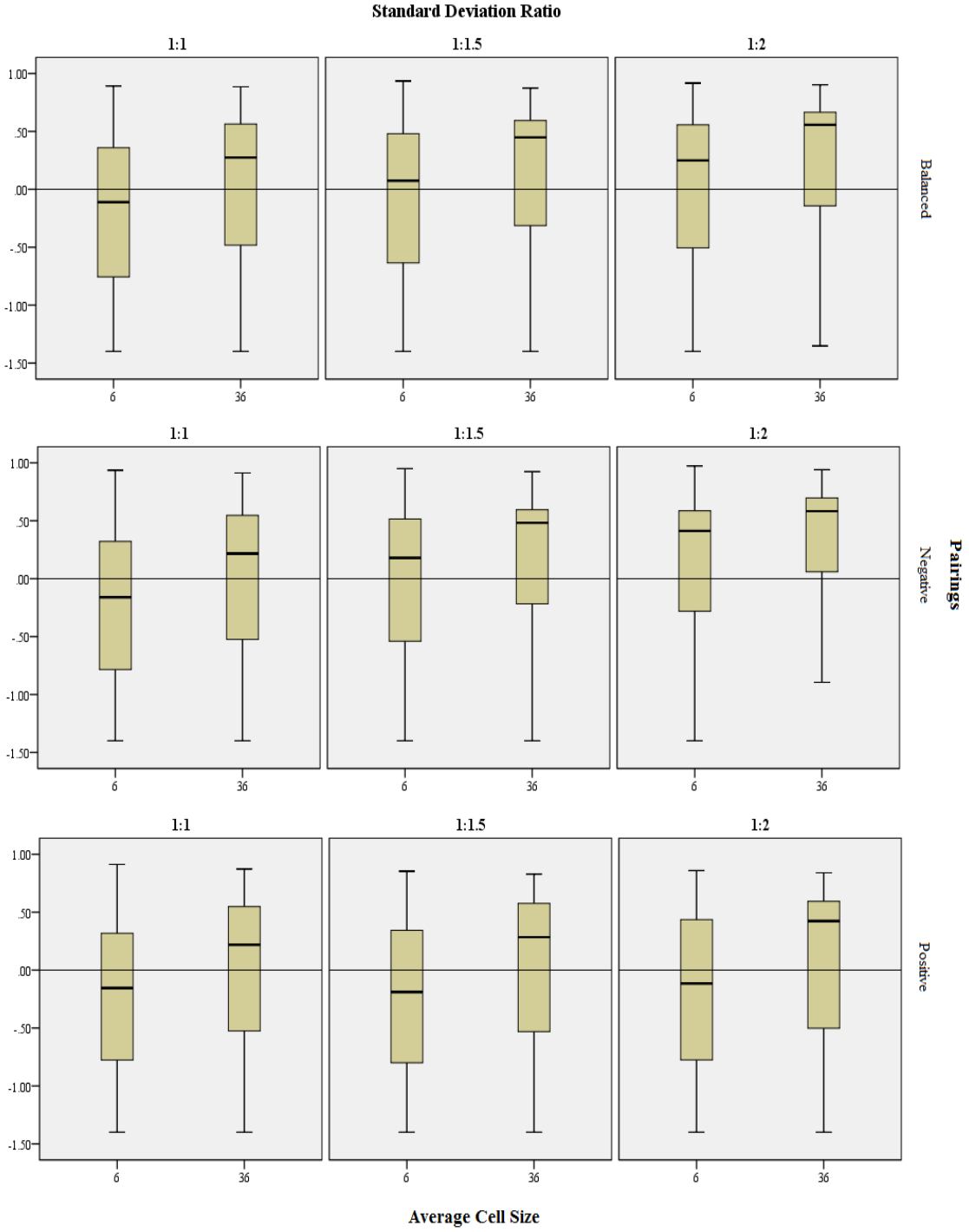

Figure 10. Box-and-Whisker plots for the sampling error bias of estimated ICC for the A-way across heterogeneity, sampling type, and with different cell size.

On a whole, as provided in Table 27, population ICC values contributed the greatest amount to the total variance (12.6%), followed by average cell size (3.1%), ratio of the group standard deviations (1.7%), and the ratio of group sizes (1%). The interaction between the population ICC and the average cell sizes explained 1.3% of the total variance, the interaction between the ratio of standard deviations and the population ICCs explained 0.7% of the total variance, and the interaction between the ratio of standard deviations and the ratio of group sizes explained 0.5% of the total variance.

What Affects the Robustness of the Estimated Parameter ICCs?

The estimated parameters were fairly robust. As shown in Table 28, the ratio of standard deviations and the sizes of population ICCs each explained 0.1% of the total variance, and the effects of all other main and interaction effects were minimal. The results indicated that even though the estimated parameters might exhibit positive or negative bias, the estimates were always consistent.

To What Extent the Size of the Fixed-Effect and Whether or Not There Was An Interaction Impact the Estimation of Parameter ICCs?

The size of the fixed effect did not much affect the estimation of the parameter ICCs. Whether or not there was an interaction effect did not affect the estimation of parameter ICCs. Provided in Figure 11 are the box-and-whisker plots for the sampling error bias of estimated ICCs for the A-way across heterogeneity, sampling type, and with/without interaction. The estimates were nearly identical regardless of the presence or absence of an interaction.

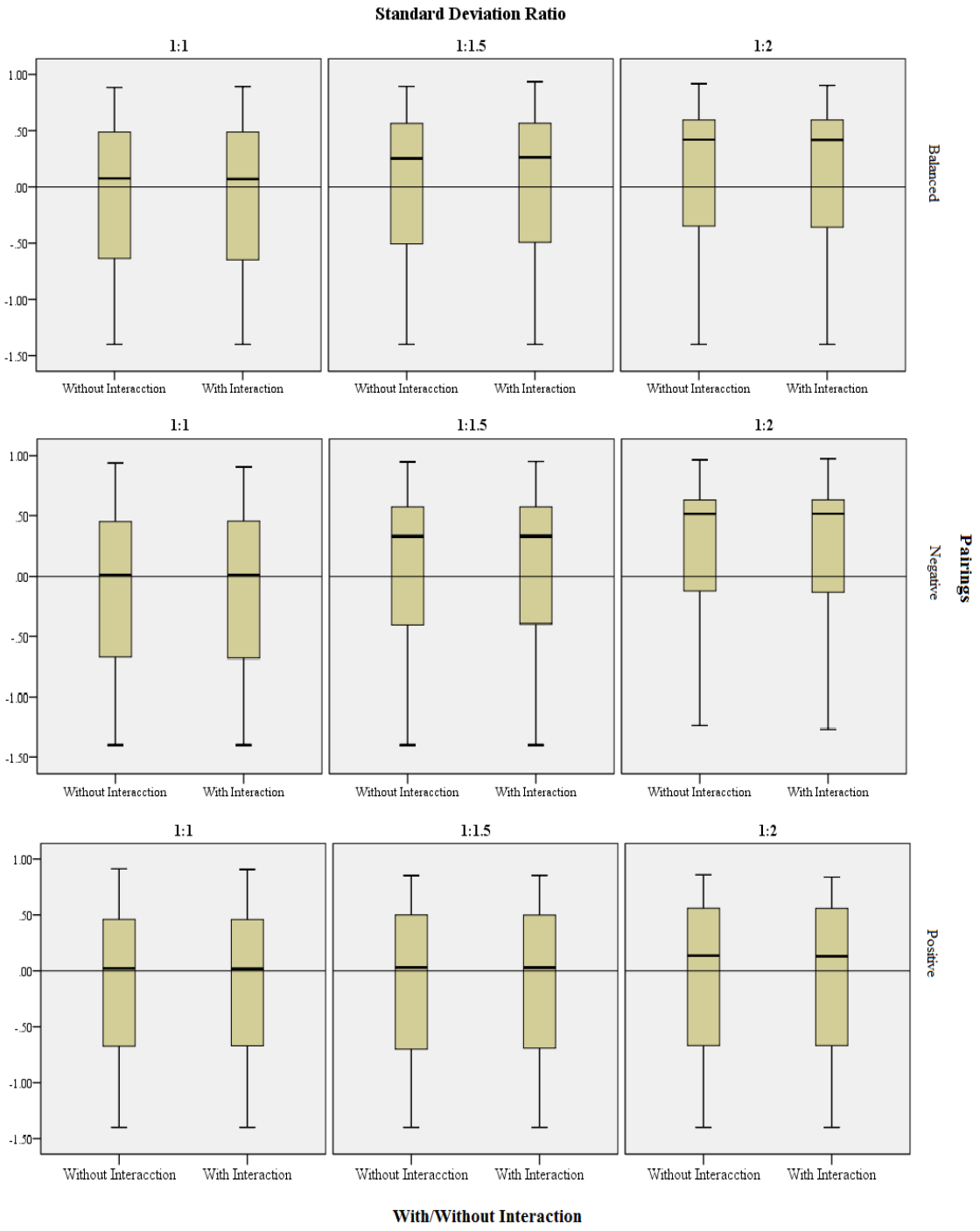

Figure 11. Box-and-Whisker plots for the sampling error bias of estimated ICC for the A-way across heterogeneity, sampling type, with/without interaction.

Summary

The ICC was defined as $\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\epsilon^2}$. During the estimation of ICCs, both the σ_α^2 and the σ_ϵ^2 need to be estimated, and both were affected by sampling error. As provided in Figure 12, the estimated σ_α^2 was composed of the true σ_α^2 and the sampling error due to the random-effect, and the estimated σ_ϵ^2 was composed of the true σ_ϵ^2 and the sampling error due to random sampling of each independent sample. And as obtained from my fixed-effects ANOVA simulation study, the sampling error due to random sampling was affected by many conditions and the interaction between these conditions, such as sample size, ratio of standard deviation, ratio of group size, pairing, etc. Similarly, the sampling error due to the random-effect was mainly affected by the number of levels randomly sampled from the population levels of effect (which is analogous to the sample size in the fixed-effects model).

Because of the complicated mechanism that affects the estimation of ICCs, each condition no longer consistently inflated or deflated the parameter estimate. Inflating or deflating was largely dependent on which sampling error (i.e., sampling error from random-effect, or sampling error from random sampling of independent sample) more inflated the true variance.

The resulting robust estimates in the present simulation study can be understood if one considers that the possible heterogeneity of variance between levels was not considered. I only considered the mean difference between each level, but variance for each level was assumed to be equal in all circumstances. But in reality, variance for each level may differ, and the experimental design will not always draw two levels to estimate

the random effect. If I had considered the heterogeneity of variance in each level, and different sample sizes for each level, the robustness of population estimation might not as great as what was observed.

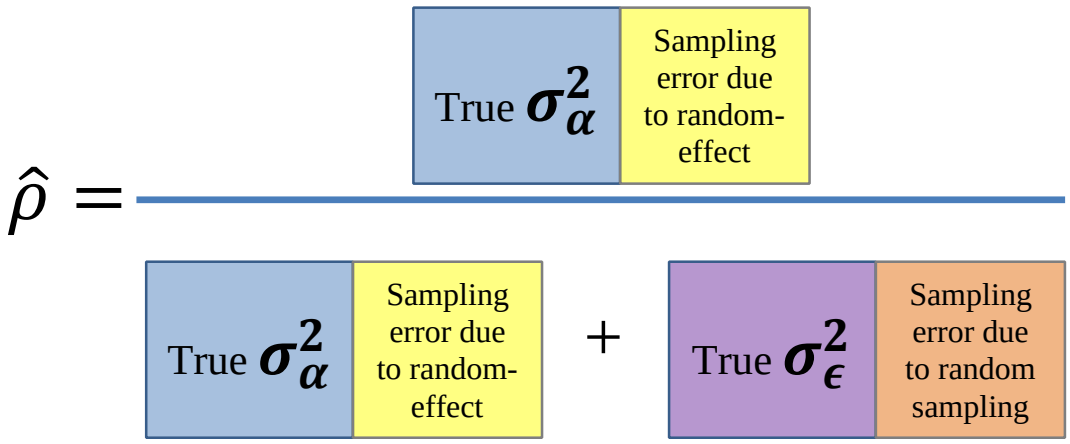

Figure 12. How does sampling error effect on the estimation of population ICCs.

Conclusion

Estimation of ICC under the framework of analysis of variance is very complicated. The present simulation study only examined the estimation of ICC in a 2 × 3 mixed-effect ANOVA design. The accuracy of estimation was mainly affected by two components: sampling error due to the random-effect and sampling error due to random sampling of a sample. Therefore, the accuracy of estimation was not only affected by the random-effect but was also affected by the factors that produce the sampling errors in each independent sample. A 2 × 3 mixed-effect ANOVA design might have produced

large sampling error variance due to the small sample size (i.e., two levels). When the sample size in each independent sample was small, two large sampling error variances might balance out, and yield a more accurate estimation. But when the sample size was large, the random-effect sampling error might have had a larger impact on the estimation of ICCs, thus yielded more biased results.

The estimation of ICC is pretty robust if researchers are only interested in a 2 × 3 mixed-effect ANOVA design. Because the sample size at the group level is fixed, there is no pairings effect as in each independent sample. The take-home message is that the ICC estimates are robust across different studies as long as the number of levels for the random effect is the same. Researchers should be cautious to utilize the ICCs' estimates when the design differs from the design investigated here.

SUMMARY AND CONCLUSIONS

As newer data analytic techniques have been introduced, the dominance of ANOVA in educational and psychological quantitative research has markedly diminished since the 70's in Education and 80s in Psychology (Skidmore & Thompson, 2010). Still, ANOVA remains a frequently used quantitative data analytic technique in education and psychology. The persistent use of ANOVA merits further investigation of such issues as misuse and misinterpretation of ANOVA. Issues, such as ignoring ANOVA assumptions and effect sizes, and relying too much on *p*-values, have been pointed out by methodologists such as Kesselman and his colleagues (1998) decades ago. The first study in this dissertation investigated ANOVA applications in three APA journals in 2012. Results revealed that nowadays, researchers that use the ANOVA technique still tend to ignore the prerequisite assumptions, but the reporting of effect sizes are on the increase. The results coincide with the trend of increasingly emphasizing the reporting of effect sizes by APA journals.

The second study further explored how violations of ANOVA assumptions impacted the estimation of fixed-effect ANOVA effect sizes. The results generated the following conclusions about multi-way ANOVA: (1) classical forms of effect sizes (i.e., $\hat{\eta}^2$, $\hat{\varepsilon}^2$, and $\hat{\omega}^2$) are more robust than partial alternative forms (i.e., $\hat{\eta}_p^2$, $\hat{\varepsilon}_p^2$, and $\hat{\omega}_p^2$); (2) $\hat{\varepsilon}^2$ and $\hat{\omega}^2$ are not always a better estimate of ANOVA effect sizes, $\hat{\eta}^2$ can offset the deflation of estimation due to negative pairing of unequal group sizes and group standard deviations, and sometimes yield more accurate estimate than $\hat{\varepsilon}^2$ and $\hat{\omega}^2$; (3) When the

number of ways increase, the effect of Cohen's *f* decreases but the effect of group size ratios increases; and (4) things that could affect the accuracy of estimation were sample size, ratio of group size, ratio of standard deviation, and the interaction between those main factors.

And the third study further explored the estimation of mixed-effect ANOVA effect sizes. Results revealed an even more complicated situation. In the mixed-effect ANOVA, there were two types of sampling errors: sampling error due to random-effects and sampling error due to random sampling of a sample. Additionally, each sampling error component could be affected by sample size, ratio of group size, ratio of group standard deviation, and the interaction between those main factors. In the ANOVA experimental design, researchers usually do not consider possible differences in standard deviation between levels, but when models are different (e.g., 2×2 vs. 3×3), the sample sizes for random ways are different. Therefore, although the accuracy of estimation of random-effect ANOVA effect sizes is affected by different conditions, the estimation itself is pretty robust. However, readers are cautioned to compare these estimations across different models because when the number of ways randomly sampled from the population ways is different, the estimation may be biased to a different extent.

REFERENCES

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: American Psychological Association.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *The Journal of Wildlife Management, 64,* 912-923.

Bangert, A. W., & Baumberger, J. P. (2005). Research and statistical techniques used in the journal of counseling & development: 1990–2001. *Journal of Counseling & Development, 83,* 480-487.

Bathke, A. (2004). The ANOVA *F* test can still be used in some balanced designs with unequal variances and nonnormal data. *Journal of Statistical Planning and Inference, 126,* 413-422.

Baumgartner, T. A., Jackson, A. S., Mahar, M. T., & Rowe, D. A. (2003). *Measurement for evaluation in physical education and exercise science* (7th ed.). Boston, MA: McGraw-Hill.

Baumberger, J. P., & Bangert, A. W. (1996). Research designs and statistical techniques used in the journal of learning disabilities, 1989-1993. *Journal of Learning Disabilities, 29*, 313-316.

Bennington, C. C., & Thayne, W. V. (1994). Use and misuse of mixed model analysis of variance in ecological studies. *Ecology, 75*, 717-722.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349–381). San Francisco, CA: Jossey-Bass.

Boswell, J. F., McAleavey, A. A., Castonguay, L. G., Hayes, J. A., & Locke, B. D. (2012). Previous mental health service utilization and change in clients' depressive symptoms. *Journal of Counseling Psychology*, 59, 368-378.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems I: Effects of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, 25, 290-302.

Cardinal, R. N., & Aitken, M. R. (2013). *ANOVA for the behavioral sciences researcher.* Mahwah, NJ: Erlbaum.

Cohen, J. (1965). *Some statistical issues in psychological research.* In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95-121). New York, NY: McGraw-Hill.

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement, 33,* 107-112.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Cohen, J. (1994). The Earth is round (p < . 05). *American Psychologist, 49*, 997-1003.

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement, 33*, 107-112.

Cook, C. (2000, January). *A review of intraclass correlation*. Paper presented at the Annual Meeting of the Southwest Educational Research Association, Texas A&M University, Dallas, TX.

Daniel, L. G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools, 5*, 23-32.

David, H. A. (1995). First (?) occurrence of common terms in mathematical statistics. *The American Statistician, 49,* 121. doi:10.2307/2684625

Dixon, M. A., & Cunningham, G. B. (2006). Data aggregation in multilevel analysis: A review of conceptual and statistical issues. *Measurement in Physical Education and Exercise Science, 10*, 85-107.

Edgington, E. S. (1964). A tabulation of inferential statistics used in psychology journals. *American Psychologist, 19,* 202-203.

Edgington, E. S. (1974). A new tabulation of statistical procedures used in APA journals. *American Psychologist, 29,* 25-26.

Fan, X., Felsovalyi, A., Sivo, S. A., & Keenan, S. C. (2002). *SAS for Monte Carlo studies: A guide for quantitative researchers.* Cary, NC: SAS Institute.

Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed-and random-effects effect sizes. *Educational and Psychological Measurement, 61*, 575-604.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh, Scotland: Oliver and Boyd.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly, 52,* 647-674.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42,* 237-288.

Goodwin, L. D., & Goodwin, W. L. (1985). An analysis of statistical techniques used in the journal of educational psychology, 1979-1983. *Educational Psychologist, 20,* 13-21.

Griffin, D. G., & Gonzalez, R. (1995). Correlational analysis of dyad-level data in the exchangeable case. *Psychological Bulletin, 118,* 430-439.

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications.* New York, NY: Routledge.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, 52. 15-24.

Haggard, E. A. (1958). *Intraclass correlation and the analysis of variance*. New York, NY: Dryden Press.

Harris, J. A. (1913). On the calculation of intra-class and inter-class coefficients of correlation from class moments when the number of possible combinations is large. *Biometrika, 9,* 446-472.

Harwell, M. R. (1991). Using randomization tests when errors are unequally correlated. *Computational Statistics & Data Analysis, 11,* 75-85.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one-and two-factor fixed effects ANOVA cases. *Journal of Educational and Behavioral Statistics, 17,* 315-339.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin, 93,* 388-395.

Hedges, L. V., & Olkin, I. (1985). *Statistical method for meta-analysis.* Orlando, FL: Academic.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis.* Hoboken, NJ: Wiley.

Hester, Y. C.(2000). *An analysis of the use and misuse of ANOVA.* Doctoral dissertation, Texas A&M University. Retrieved from Pro Quest LLC.(UMI 9994257).

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah, NJ: Erlbaum.

Hsu, P. L. (1938). Contribution to the theory of "student's" west as applied to the problem of two samples. *Statistical Research Memoirs, 2,* 1-24.

Hurst, J. J. (1996). *Methodological issues in research on couples.* Dissertation Abstracts International: Section B: The Sciences and Engineering.

Kane, M. (2002). Inferences about variance components and reliability-generalizability coefficients in the absence of random sampling. *Journal of Educational Measurement, 39*, 165-181.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17,* 137-152.

Kerlinger, F. N. (1973). *Foundations of behavioral research: Educational, psychological and sociological inquiry*. New York, NY: Holt, Rinehart and Winston.

Keselman, H. (1975). A Monte Carlo investigation of three estimates of treatment magnitude: Epsilon squared, eta squared, and omega squared. *Canadian Psychological Review, 16*, 44-48.

Keselman, H., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., . . . Keselman, J. C. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350-386.

Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in AERJ and JCP articles from 1988 to 1997: A methodological review. *The Journal of Experimental Education, 69*, 280-309.

Klein, K. J., Bliese, P. D., Kozlowski, S. W. J., Dansereau, F., Gavin, M. B., Griffin, M. A., Hofmann, D. A., James, L. R., Yammaeino, F. J., & Bligh, M. C.. (2000). Multilevel analytical techniques: Commonalities, differences, and continuing questions. In K. J. Kline & S. W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 512–553). San Francisco, CA: Jossey-Bass.

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research, 28*, 612-625.

Lix, L. M., Keselman, J. C., & Keselman, H. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66*, 579-619.

Luh, W., & Guo, J. (2001). Using johnson's transformation and robust estimators with heteroscedastic test statistics: An examination of the effects of non-normality and heterogeneity in the non-orthogonal two-way ANOVA design. *British Journal of Mathematical and Statistical Psychology, 54,* 79-94.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30-46.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.). *What if there were no Significance Tests* (pp. 393-425). Mahwah, NJ: Lawrence Erlbaum.

Milligan, G. W., Wong, D. S., & Thompson, P. A. (1985). An algorithm for testing the robustness properties of two-way nonorthogonal analysis of variance. *Educational and Psychological Measurement, 45,* 607-611.

Milligan, G. W., Wong, D. S., & Thompson, P. A. (1987). Robustness properties of nonorthogonal analysis of variance. *Psychological Bulletin, 101,* 464-470.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews, 82,* 591-605.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25,* 241-286.

Paterson, L., & Goldstein, H. (1991). New statistical methods for analysing social structures: An introduction to multilevel models. *British Educational Research Journal, 17,* 387-393.

Pearson, K. (1900). Mathematical contributions to the theory of evolution. VII. on the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London.Series A, Containing Papers of a Mathematical Or Physical Character (1896-1934),* 195(262), 1-405. doi:10.1098/rsta.1900.0022

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement, 64,* 916-924.

Rico, R., Sánchez-Manzanares, M., Antino, M., & Lau, D. (2012). Bridging team faultlines by combining task role assignment and goal structure strategies. *Journal of Applied Psychology,* 97, 407.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 115-129.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shrout, P. E. (1995). Measuring the degree of consensus in personality judgments. In P. E. Shrout & S. T. Fiske (Eds.), *Personality research, methods, and theory: A festschrift honoring donald W. fiske* (pp. 79 –92). Hillsdale, NJ: Erlbaum.

Shrout, P. S., & Fleiss, J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.

Silverman, R. J. (1987). How we know what we know: A study of higher education journal articles. *The Review of Higher Education, 11,* 39-59.

Skidmore, S. T. (2009). *Effect size matters: Empirical investigations to help researchers make informed decisions on commonly used statistical techniques*. Doctoral dissertation, Texas A&M University. Available electronically from http://hdl.handle.net/1969.1/ETD-AMU-2009-12-7460.

Skidmore, S. T., & Thompson, B. (2013). Bias and precision of some classical ANOVA effect sizes when assumptions are violated. *Behavior Research Methods, 45,* 536-546.

Skidmore, S. T., & Thompson, B. (2010). Statistical techniques used in published articles: A historical review of reviews. *Educational and Psychological Measurement, 70,* 777-795.

Thompson, B. (1989). Statistical significance, result importance, and result generalizability: Three noteworthy but somewhat different issues. *Measurement and Evaluation in Counseling and Development, 22,* 2-6.

Thompson, B. (1996). Research news and comment: AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology, 9*, 165-181.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31(3)*, 25-32.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *The Journal of Experimental Education, 61*, 361-377.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York, NY: Guilford.

Thompson, B., Diamond, K. E., McWilliam, R., Snyder, P., & Snyder, S. W. (2005). Evaluating the quality of evidence from correlational research for evidence-based practice. *Exceptional Children, 71*, 181-194.

Trafimow, D. & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, 37, 1-2.

Walsh, B.D. (1996). A note on factors that attenuate the correlation coefficient and its analogs. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 4, pp. 21-32). Greenwich, CT: JAI Press.

Willson, V. L. (1980). Research techniques in" AERJ" articles: 1969 to 1978. *Educational Researcher, 9(6),* 5-10.